# Αυτόνομη Πλοήγηση Θαλάσσιας Ρομποτικής Πλατφόρμας με χρήση Μεθόδων Ενισχυτικής Μάθησης

Η Μεταπτυχιακή Εργασία Εξειδίκευσης

υποβάλλεται στην ορισθείσα

από τη Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Μηχανικών Η/Υ και Πληροφορικής

Εξεταστική Επιτροπή

από τον

Κωνσταντίνο Τζιορτζιώτη

ως μέρος των υποχρεώσεων για την απόκτηση του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ

ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ - ΕΦΑΡΜΟΓΕΣ

Πανεπιστήμιο Ιωαννίνων

Ιούλιος 2016

# Αφιερωση

Στην οικογένειά μου

# Ευχαριστιες

Αρχικά, θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες στον επιβλέποντα καθηγητή της μεταπτυχιακής αυτής διατριβής, Αναπληρωτή Καθηγητή κ. Κωνσταντίνο Μπλεκα για την εμπιστοσύνη που επέδειξε στο πρόσωπό μου αναθέτοντας μου αυτή την εργασία, την πολύτιμη βοήθειά του, την καθοδήγησή του αλλά και για τις γνώσεις που μου μετέδωσε καθ'όλη τη διάρκειά της.

Επίσης, νιώθω την ανάγκη να ευχαριστήσω τον Επίκουρο Καθηγητή, κ. Βλάχο Κωνσταντίνο για την αποτελεσματική συνεργασία και συμβολή του στην ολοκλήρωση της παρούσας διατριβής.

Επιπλέον, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στην οικογενειά μου για την αμέριστη ηθική και οικονομική υποστήριξη που μου παρείχαν σε όλη την διάρκεια τόσο των προπτυχιακών αλλά όσο και των μεταπτυχιακών σπουδών μου.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ σε όλους τους δικούς μου ανθρώπους για την κατανόηση και την υπομονή που επέδειξαν όλο αυτό το διάστημα.

# Περιεχομενα

# Κατάλογος Σχημάτων

# Κατάλογος Αλγορίθμων

# Περιληψη

Κωνσταντίνος Τζιορτζιώτης, Μ.Δ.Ε. στην Πληροφορική, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιούλιος 2016.
Αυτόνομη Πλοήγηση Θαλάσσιας Ρομποτικής Πλατφόρμας με χρήση Μεθόδων Ενισχυτικής Μάθησης.
Επιβλέπων: Κωνσταντίνος Μπλέκας, Αναπληρωτής Καθηγητής.

Η παρούσα εργασία πραγματεύεται την αυτόνομη πλοήγηση μιας θαλάσσιας ρομποτικής πλατφόρμας - *Delta Berenike* - μέσω μεθόδων ενισχυτικής μάθησης (*reinforcement learning*), δηλ. της βέλτιστης κίνησης της πλατφόρμας με στόχο τον εντοπισμό μιας συγκεκριμένης θέσης-στόχου και με ταυτόχρονη αποφυγή εμποδίων και συγκρούσεων. Μερικές βασικές ιδιαιτερότητες της συγκεκριμένης θαλάσσιας πλατφόρμας που σχετίζονται άμεσα με το σύστημα ελέγχου αποτελουν οι διαταραχές εξαιτίας του υδροδυναμικού μοντέλου και του σύνθετου δυναμικού μοντέλου των επενεργητών, όπως επίσης τα σφάλματα που προέρχονται από την ανατροφοδότηση των μεταβλητών κατάστασης (τρέχουσα θέση, προσανατολισμός και ταχύτητα) από τους αισθητήρες κίνησης. Το πλαίσιο της ενισχυτικής μάθησης προσεγγίζει το πρόβλημα της πλοήγησης ως ένα πρόβλημα ανακάλυψης της βέλτιστης πολιτικής ενός πράκτορα (*agent*) ο οποίος κινείται σε ένα στοχαστικό Μαρκοβιανό χώρο καταστάσεων. Κατά την διάρκεια της αλληλεπίδρασης του πράκτορα με το περιβάλλον σημαντικό ρόλο στη διαδικασία μάθησης αποτελεί η συνάρτηση ανταμοιβής (*reward function*), η οποία καθορίζει την μορφή απεικόνισης του χώρου καταστάσεων με τις ενέργειες (συνάρτηση αξίας *action value function*). Η συνήθης διαδικασία είναι ότι η συνάρτηση ανταμοιβής είναι εκ των προτέρων γνωστή με βάση την εμπειρία του προβλήματος ελέγχου. Γενικά το πρόβλημα της εκτίμησης της συνάρτησης ανταμοιβής ορίζεται ως ένα πρόβλημα αντίστροφης ενισχυτικής μάθησης (*inverse reinforcemenet learning*).

Στην εργασία αυτή προτείνεται ένα πλαίσιο ενισχυτικής μάθησης για τον έλεγχο της θαλάσσιας πλατφόρμας, το οποίο εκτιμά τη βέλτιστη πολιτική και ταυτόχρονα τη συνάρτηση ανταμοιβής. Η διαδικασία μάθησης είναι on-line και επικεντρώνεται στο γραμμικό μαθηματικό μοντέλο (*linear model*) για την περιγραφή των συναρτήσεων αξιών και ανταμοιβών, χρησιμοποιώντας ένα περιγραφικό χώρο καταστασεων μέσω κατάλληλων συναρτήσεων βάσης (*basis functions*). Η προτεινόμενη μέθοδος αξιολογήθηκε πειραματικά σε προσομοιωμένα θαλάσσια στοχαστικά περιβάλλοντα στα οποία επιδρούν διάφορες μορφές περιβαλλοντικών διαταραχών, όπως ο άνεμος, τα θαλάσσια ρεύματα καθώς και τα κύματα. Τέλος, πραγματοποιήθηκε η σύγκριση της μεθόδου με δύο γνωστές τεχνικές ενισχυτικής μάθησης, τον αλγόριθμο *Q-Learning* και τον *LSPI*.

# Abstract

Konstantinos Tziortziotis, M.Sc. in Computer Science, Department of Computer Science and Engineering, University of Ioannina, Greece, July 2016.
Autonomous navigation of an over-actuated marine platform using reinforcement learning.
Advisor: Konstantinos Blekas, Associate Professor.

The current diploma thesis examines the autonomous navigation of an over-actuated marine platform "Delta Berenike" using reinforcement learning methods. RL aims at finding an optimum route through obstacles in order to identify a specific target. The marine platform is related with certain peculiarities due to hydrodynamic model of the actuators and complex dynamic model,as well as errors from feedback of state variables (current position, orientation and speed) from motion sensors. The RL framework considers navigation problem as a problem of finding the optimal policy of an agent which moves in a stochastic Markov state space. During the agent's interaction with the environment, reward function plays an important role in the learning process, determining the display format of the state space with the actions. More specifically, reward function is known in advance and is based on the control problem experience.Generally, the problem of the reward function estimation is defined as an inverse reinforcement learning problem.

In this study we propose a reinforcement learning framework which controls marine platform and estimates the optimum policy as well as the reward function. The learning process can be implemented as an on line learning algorithm and is focused on linear model in order to describe the value and rewards functions, using a descriptive situation space through appropriate basis function. We have studied the performance of the proposed method using two simulated environments. The environment includes several environmental disturbances such as wind, sea currents and waves. Finally, emphasis was given in the comparison of the method through

two known reinforcement learning techniques, the Q-Learning and LSPI algorithm. The results are promising.

# Κεφαλαιο 1

# Εισαγωγη



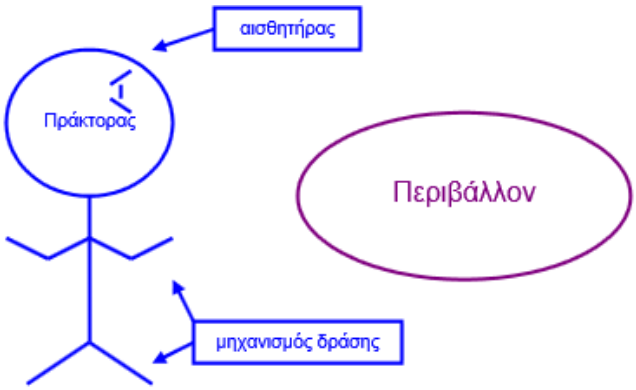
## 1.1 Τεχνητή Νοημοσύνη-ΤΝ (Artificial Intelligence-AI)

Από τα πρώτα στάδια της ιστορίας ο άνθρωπος επιδίωξε την δημιουργία μηχανισμών που θα είχαν την δυνατότητα να σκέφτονται, να αποφασίσουν και να ενεργούν όπως ο ίδιος. Αυτός ο διακαής πόθος λοιπόν να γίνει ο δημιουργός ενός τεχνητού έργου που θα παρουσιάζει μια μορφή ευφυΐας οδήγησε στην διαμόρφωση σήμερα της επιστήμης της Τεχνητής Νοημοσύνης. Η Τεχνητή Νοημοσύνη (ΤΝ) (*Artificial Intelligence - AI*) είναι η μελέτη και δημιουργία προγραμμάτων Η/Υ που σκοπό έχουν να συμπεριφέρονται "έξυπνα". Ο όρος "έξυπνα" δεν έχει σαφή ερμηνεία, αλλά για τον χώρο της τεχνητής νοημοσύνης μπορεί να ερμηνευτεί σαν: ικανότητα να επιλύουν προβλήματα, να μαθαίνουν από προηγούμενες εκτελέσεις, να καταλαβαίνουν δύσκολες καταστάσεις και δεδομένα, και να αντιλαμβάνονται διαφορές και ομοιότητες μεταξύ καταστάσεων. Αποτελεί σημείο τομής μεταξύ πολλαπλών επιστημών όπως της πληροφορικής, της ψυχολογίας, της φιλοσοφίας, της νευρολογίας, της γλωσσολογίας

και της επιστήμης μηχανικών, με στόχο τη σύνθεση ευφυούς συμπεριφοράς. Κατά τη διάρκεια των τελευταίων δύο δεκαετιών η τεχνητή νοημοσύνη έχει παρουσιάσει αλματώδη ανάπτυξη και υπάρχει σαφώς ένα σημαντικό ερευνητικό ενδιαφέρον για την κατασκευή ευφυών πρακτόρων για ψηφιακά παιχνίδια που μπορούν να προσαρμόσουν την συμπεριφορά των παικτών καθώς και για την πλοήγηση ρομποτικών κατασκευών σε περιβάλλοντα που αλλάζουν δυναμικά.

## 1.2 Η έννοια του πράκτορα (agent)

Ο όρος πράκτορας (*agent*) έχει πολλούς ορισμούς και έννοιες. Κάτω από την ομπρέλα της έννοιας "πράκτορας" έχει αναπτυχθεί ένα ετερογενές πεδίο έρευνας. Θα μπορούσαμε να πούμε ότι ένας πράκτορας είναι μια οντότητα που αντιλαμβάνεται το περιβάλλον (*environment*) μέσα στο οποίο βρίσκεται με τη βοήθεια αισθητήρων (*sensors*), κάνει συλλογισμούς για το περιβάλλον και δρα πάνω σε αυτό με τη βοήθεια μηχανισμών δράσης (*effectors*) για την επίτευξη κάποιον στόχων (goals)(Russell & Norvig, 1995). Πράκτορας εναλλακτικά είναι ενα κομμάτι λογισμικού και/ή υλικού το οποίο είναι ικανό να δρα προκειμένου να εκτελέσει κάποιες εργασίες εκ μέρους των χρηστών του(Nwana,1996)




Σχήμα 1.1: Αναπαράσταση ενός πράκτορα

Ένας ευφυής πράκτορας έχει την ικανότητα εκτέλεσης μιας ευέλικτης αυτόνομης ενέργειας προκειμένου να επιτύχει τους στόχους σχεδιασμού του. Ο όρος ευέλικτη ενέργεια εμπεριέχει τα τρία ακόλουθα χαρακτηριστικά:

- αντιδραστικότητα
- προνοητικότητα
- κοινωνικότητα.

Πιο συγκεκριμένα ευέλικτη ενέργεια είναι η ικανότητα του πράκτορα να ακολουθεί το περιβάλλον του και να αντιδρά εντος κάποιων χρονικών περιθωρίων στις αλλαγές που παρατηρούνται σε αυτό, να εμφανίζει συμπεριφορά που κατευθύνεται από τους στόχους και να αλληλεπιδρά με τους άλλους πράκτορες προκειμένου να ικανοποιήσουν τους στόχους σχεδιασμού τους.

## 1.3 Αυτόνομη Πλοήγηση Ρομποτικών Συστημάτων

Ρομποτικός πράκτορας ονομάζεται οποιαδήποτε ευφυής μηχανική συσκευή που μπορεί να υποκαθιστά τον άνθρωπο σε διάφορες εργασίες. Ένα ρομπότ μπορεί να δράσει αυτόνομα σε ένα οποιοδήποτε περιβάλλον και είναι σε θέση να αντιλαμβάνεται, να σκέφτεται και να ενεργεί.


Σχήμα 1.2: Μη επανδρωμένο θαλάσσιο ρομποτ

Τα ρομπότ μπορούν να χρησιμοποιηθούν ώστε να κάνουν εργασίες οι οποίες είτε είναι δύσκολες ή επικίνδυνες για να γίνουν απευθείας από έναν άνθρωπο. Ωστόσο είναι κατάλληλα εργαλεία για την διευρεύνηση προβλημάτων τεχνητής νοημοσύνης που περιλαμβάνουν την αποφυγή εμποδίων ή τη σχεδίαση μοναπατιού. Ένα από τα χαρακτηριστικά των ρομποτικών συστημάτων είναι η πολυπλοκότητα των περιβάλλοντων μέσα στα οποία κινούνται.

Η πλοήγηση είναι μια ουσιώδη ικανότητα ενός αυτόνομου κινούμενου ρομποτικού συστήματος. Του δίνει την δυνατότητα να παραμένει σε λειτουργία αποφεύγοντας τη σύγκρουσή του με εμπόδια ενώ του επιτρέπει επίσης να φθάσει σε συγκεκριμένες περιοχές ενός άγνωστου περιβάλλοντος που σχετίζονται με κάποιο συγκεκριμένο έργο που πρέπει να φέρει εις πέρας. Πρακτικά το ρομπότ δεν μπορεί να ανακαλύψει άμεσα ένα μονοπάτι από κάποιο αρχικό σημείο προς ένα προορισμό. Για το λόγο αυτό θα πρέπει να χρησιμοποιηθούν τεχνικές εύρεσης μονοπατιού που συνεπάγονται τη μετάβαση από μια αρχική θέση σε κάποιο προορισμό ενώ ταυτόχρονα ελαχιστοποιούν κάποιο κόστος. Η πλοήγηση σχετίζεται με τρία τμήματα:

1. την εύρεση του ρομπότ στο χώρο (*localization*)
2. τη χαρτογράφηση του χώρου (*mapping*)
3. το σχεδιασμό του μονοπατιού (*path planning*)

Ο σχεδιασμός του μονοπατιού αποτελεί το σημαντικότερο βήμα της διαδικασίας της πλοήγησης καθώς με τη διαδικασία αυτή το ρομπότ βρίσκει το βέλτιστο μονοπάτι που πρέπει να ακολουθήσει προκειμένου να φθάσει στο στόχο αποφεύγοντας εμπόδια, επικίνδυνες περιοχές κτλ. στην πορεία του. Αρκετές προσεγγίσεις έχουν προταθεί για το πρόβλημα της σχεδίασης της κίνησης ενός ρομποτικού συστήματος μεσα σε ένα περιβάλλον. Αυτές οι προτάσεις περιλαμβάνουν αλγορίθμους είτε *off-line*, δηλαδή που παράγουν εκ των προτέρων ένα μονοπάτι για ένα ήδη γνωστό στατικό περιβάλλον είτε απευθείας, *on-line*, έχοντας την δυνατότητα εύρεσης του καινούργιου μονοπατιού εξαιτίας κάποιας αλλαγής που συμβαίνει στο περιβάλλον.

## 1.4 Αντικείμενο της Διατριβής

Στην παρούσα διατριβή εστιάζουμε στην μελέτη μεθόδων ενισχυτικής μάθησης για τη δημιουργία ενός ευφυή πράκτορα που θα πλοήγει τη θαλάσσια ρομποτική πλατφόρμα *Delta Berenike*. Η Μηχανική Μάθηση (*Machine Learning*) ασχολείται με την ανάπτυξη μεθόδων που επιτρέπουν σε ένα υπολογιστικό σύστημα να βελτιώνει τη συμπεριφορά του μέσω της μάθησης και της προσαρμογής. Η έννοια της εκπαίδευσης εντοπίζεται κυρίως πάνω στην εύρεση κατάλληλων παραμέτρων των μοντέλων ώστε να πετύχουμε καλύτερο ταίριασμα με το εκάστοτε πρόβλημα στο χώρο αναζήτησης. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να οργανωθούν σε τρεις

κατηγορίες σύμφωνα με το πρόβλημα που αντιμετωπίζουν, τη μάθηση με επίβλεψη (*supervised learning*), τη μάθηση χωρίς επίβλεψη (*unsupervised learning*) και η ενισχυτική μάθηση (*reinforcement learning*).

Η ενισχυτική μάθηση αντιμετωπίζει το πρόβλημα της μάθησης μιας βέλτιστης συμπεριφοράς από εναν πράκτορα ο οποίος ενεργεί σε ένα περιβάλλον. Ο πράκτορας λαμβάνει μέσω των αισθητηρίων οργάνων του μια αναπαράσταση της τρέχουσας κατάστασης του περιβάλλοντος και ενεργεί σύμφωνα με κάποια πολιτική. Το περιβάλλον αποκρίνεται στις ενέργειές του, παρέχοντας του αριθμητικές ανταμοιβές και παρουσιάζοντας του νέες καταστάσεις. Στόχος του πράκτορα είναι η μεγιστοποίηση της συνολικής ανταμοιβής που λαμβάνει από το περιβάλλον. Κάθε χρονική στιγμή απεικονίζει τις καταστάσεις σε πιθανότητες επιλογής όλων των δυνατών ενεργειών κι αυτή η απεικονιση καλείται πολιτική.

Αντικείμενο της διατριβής αποτελεί η πλοήγηση μιας θαλάσσιας ρομποτικής πλατφόρμας η οποία εκκινεί από τυχαίες θέσεις, αποφεύγει πιθανά εμπόδια σε ένα άγνωστο περιβάλλον και επιδιώκει να φθάνει σε μια συγκεκριμένη θέση. Αυτό επιτυγχάνεται με τη χρήση μεθόδων ενισχυτικής μάθησης. Κατά τη διάρκεια της πλοήγησής της πάνω της επιδρούν περιβαλλοντικές διαταραχές οπως ο άνεμος, τα θαλάσσια ρεύματα και τα κύματα, οι οποίες δυσκολεύουν την κίνησής της. Στις μεθόδους ενισχυτικής μάθησης η συνάρτηση ανταμοιβής είναι αυτή που ορίζει το στόχο του πράκτορα και είναι καθορισμένη. Το πρόβλημα της εύρεσης της συνάρτησης ανταμοιβής καλείται να αντιμετωπίσει η αντίστροφη ενισχυτική μάθηση η οποία δοθετος μιας βέλτιστης πολιτικης στοχεύει στην εκτίμηση της συνάρτησης ανταμοιβής. Η προσφορά της παρούσας διατριβής είναι η εφαρμογή κι αξιολόγηση μιας μεθόδου για την ταυτόχρονη εύρεση τόσο της πολιτικής αλλά και των ανταμοιβών που λαμβανει ο πράκτορας.

## 1.5 Δομή της Διατριβής

Η παρούσα διατριβή περιέχει πέντε κεφάλαια. Στο κεφάλαιο 2 παρουσιάζεται μια εκτενής επισκόπηση του πεδίου της ενισχυτικής μάθησης. Στο κεφάλαιο 3 περιγράφονται λεπτομερώς πληροφορίες για την θαλάσσια ρομποτική πλατφόρμα. Συγκεκριμένα παρουσιάζονται αρχικά τα φυσικά χαρακτηριστικά της και στη συνέχεια το μοντέλο κίνησής της. Επίσης, περιγράφεται και η μοντελοποίηση των περιβαλ-

λοντικών διαταραχών που ενσωματώσαμε στην προσομοίωση της ρομποτικής κατασκευής. Στην συνέχεια, παρουσιάζεται αναλυτικά το αλγοριθμικό σχήμα για την ταυτόχρονη εύρεση, τόσο της πολιτικής όσο και της συνάρτησης των ανταμοιβών. Τέλος, στο κεφάλαιο 5 υλοποιούνται και αξιολογουνται ορισμένες από τις κυριότερες μεθόδους ενισχυτικής μάθησης που περιγράφονται στην παρούσα διατριβή σε δύο διαφορετικά προσομοιώμενα περιβάλλοντα.

# Κεφαλαιο 2

# Ενισχυτικη Μαθηση

---

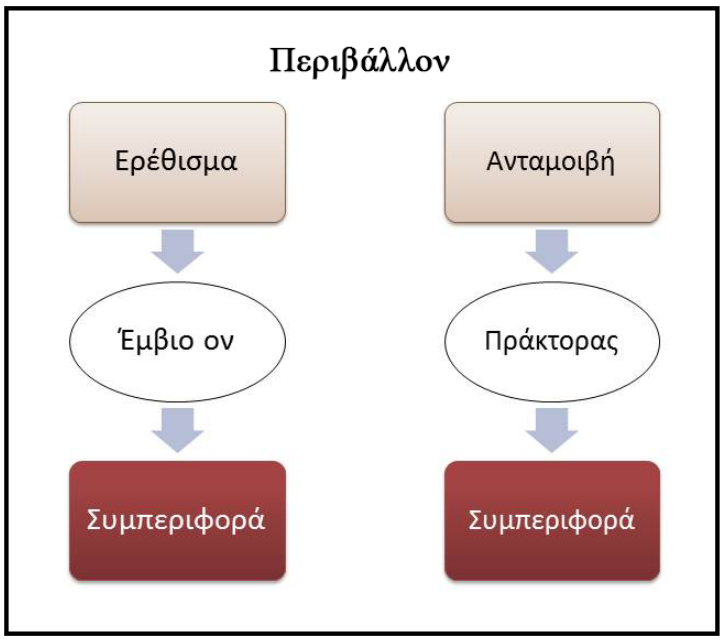


---

## 2.1 Εισαγωγή

Η Ενισχυτική Μάθηση (*Reinforcement Learning*) αποτελεί μια διαδικασία μάθησης η οποία είναι εμπευσμένη από θεωρίες της ψυχολογίας που εξηγούν με ποιο τρόπο τα έμβια όντα μαθαίνουν διάφορες συμπεριφορές. Στο Σχήμα 2.1 φαίνεται η σχέση της ενισχυτικής μάθησης μεταξύ ενός τεχνητού υπολογιστικού συστήματος και ενός ζωντανού οργανισμού.

Η ενισχυτική μάθηση έχει να κάνει με το πως ένας πράκτορας μπορεί να μάθει μια συγκεκριμένη συμπεριφορά μέσω της αλληλεπιδρασής του με το περιβάλλον




Σχήμα 2.1: Απεικόνιση της διαδικασίας μάθησης μεταξύ ενός ζωντανού οργανισμού κι ενός τεχνητού υπολογιστικού συστήματος (πράκτορα)

στο οποίο ενεργεί. Σκοπός του είναι να μεγιστοποιήσει ένα σήμα ανταμοιβής, το οποίο λαμβάνει από το περιβάλλον ως αποτέλεσμα των ενεργειών του. Ο πράκτορας δεν λαμβάνει κάποια άλλη απόκριση από το περιβάλλον εκτός από τις ανταμοιβές, γεγονός που σημαίνει ότι θα πρέπει να μάθει μόνος του πως να συμπεριφέρεται με περιορισμένη πληροφορία. Αυτή είναι και η σημαντικότερη διαφορά της ενισχυτικής μάθησης από τη μάθηση με επίβλεψη (*supervised learning*). Στην επιβλεπόμενη μάθηση που χρησιμοποιείται σε πολλούς ερευνετικούς τομείς όπως η μηχανική μάθηση, τα νευρωνικά δίκτυα και η αναγνώριση προτύπων, υπάρχει ένας πεπειραμένος εξωτερικός επιβλέπων, ο οποίος παρέχει για κάθε ενέργεια του πράκτορα την επιθυμητή ενέργεια. Η αριθμητική ανταμοιβή που λαμβάνεται ως μόνη πληροφορία κατά την διαδικασία της μάθησης καθιστά αυτά τα προβλήματα δυσκολότερα.

Ένα απλό παράδειγμα για να γίνει κατανοητή η έννοια της ενισχυτικής μάθησης είναι η περιπέτεια ενος σκύλου μέσα σε έναν λαβύρινθο. Ως ενέργειες του σκύλου θεωρούμε τις κινήσεις του και ως κατάσταση την θέση στην οποία βρίσκεται κάθε χρονική στιγμή, η οποία αλλάζει συνεχώς σαν συνέπεια των κινήσεών του. Ο σκύλος λαμβάνει μηδενική ανταμοιβή όσο βρίσκεται μέσα στον λαμβύρινθο και περιπλανιέται, αλλά ανταμοίβεται με ένα μεγάλο κόκαλο όταν βγαίνει απ' αυτόν. Σε αυτό το απλό παράδειγμα μπορούμε να διακρίνουμε τα βασικά χαρακτηριστικά της ενισχυτικής μάθησης τα οποία είναι η αναζήτηση με δοκιμή και λάθος και η αντα-

μοιβή. Στο σκύλο δεν παρέχεται καμία οδηγία ως προς το ποια θα ήταν η καλύτερη επιλογή σε κάποια διασταύρωση στον λαβύρινθο. Κάθε φορά ενημερώνεται για την ποιότητα της επιλογής του μέσω μιας τιμής ανταμοιβής. Μόνος του καλείται να επιλέξει ποιες κινήσεις είναι οι καλύτερες. Ας υποθέσουμε τώρα ότι ο σκύλος παίρνει μια σημαντική απόφαση για την κατεύθυνση που θα ακολουθήσει σε μια διασταύρωση του λαβύρινθου η οποία βρίσκεται σχετικά κοντά στην είσοδό του και η οποία βοήθησε το σκύλο να φτάσει στην έξοδο μετά από συγκεκριμένο χρόνο. Πρέπει να επισημάνουμε ότι η θετική ανταμοιβή, αποτέλεσμα αυτής της επιλογής θα δοθεί στο σκύλο όταν εν τέλει καταφέρει να βγει από τον λαβύρινθο. Ο σκύλος-πράκτορας θα κληθεί τότε να ξεδιαλύνει ποιες αποφάσεις του κατά την περιπλάνησή του στο λαβύρινθο και σε ποιον βαθμό τον βοήθησαν στο να βρει την έξοδο και να λάβει αυτήν την ανταμοιβή.

Είναι προφανές ότι οι τεχνικές μάθησης που βασίζονται σε ένα τέτοιο μοντέλο είναι πολύ ελκυστικές. Κι αυτό διότι αν οι πράκτορες υπόκεινται σε ενισχυτική μάθηση, οι σχεδιαστές καλούνται μόνο να παράγουν την τιμή της ανταμοιβής. Η τιμή της ανταμοιβής είναι απαραίτητο να ανταποκρίνεται με ακρίβεια στο σύνολο των στόχων του συστήματος, πράγμα που τελικά αποδεικνύεται ότι δεν είναι και πολύ εύκολο.

Στη συνέχεια του κεφαλαίου, παρουσιάζονται βασικές έννοιες της ενισχυτικής μάθησης οι οποίες χρησιμοποιούνται στα επόμενα κεφάλαια. Στην ενότητα 2.2 περιγράφεται το πλαίσιο της ενισχυτικής μάθησης. Στην 2.3 περιγράφεται ο τρόπος μοντελοποίησης των προβλημάτων ενισχυτικής μάθησης και στο 2.4 οι συναρτήσεις αξίας που κατέχουν κυρίαρχο ρόλο στην εκτίμηση και την εύρεση των πολιτικών. Στην συνέχεια παρουσιάζονται μέθοδοι για την επίλυση προβλημάτων ενισχυτικής μάθησης. Έπειτα αναφέρεται το ζήτημα της ανιστάθμισης της εξερεύνησης και της εκμετάλλευσης. Τέλος, στη 2.9 παρουσιάζεται ο *LSPI* αλγόριθμος για την εκτίμηση της πολιτικής ενώ στην 2.10 η μέθοδος αντίστροφης ενισχυτικής μάθησης που αντικείμενό της είναι η εύρεση της συνάρτησης ανταμοιβής δοθέντος μιας βέλτιστης πολιτικής.

### 2.1.1 Βασικά Στοιχεία της Ενισχυτικής Μάθησης

Στο σημείο αυτό κρίνεται αναγκαίο να δωθούν ορισμένες βασικές έννοιες της ενισχυτικής μάθησης.

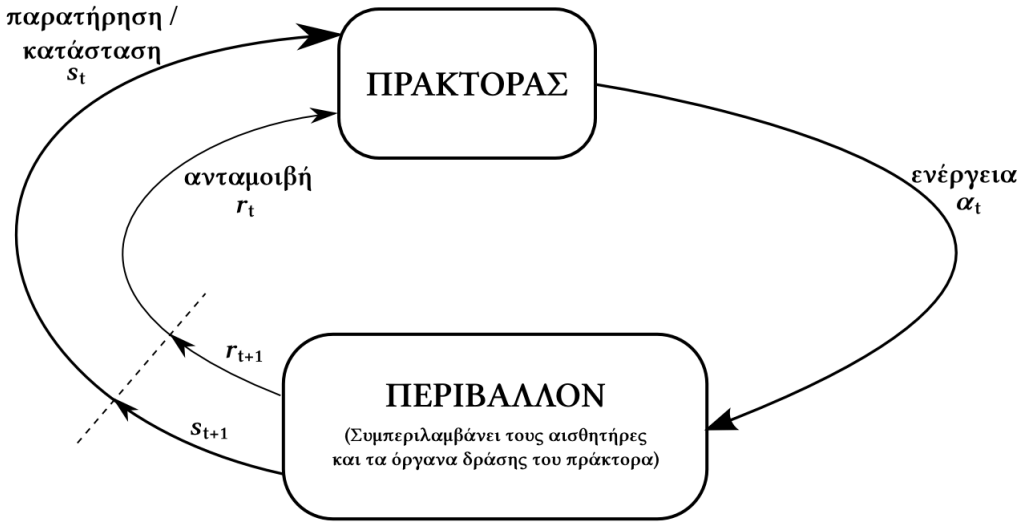
Ο πράκτορας (*agent*) είναι οποιαδήποτε οντότητα που μπορεί να αντιλαμβάνεται το περιβάλλον μέσα στο οποίο βρίσκεται μέσω αισθητήρων (*sensors*) και να ενεργεί αυτόνομα σε αυτό μέσω μηχανισμών δράσης (*effectors*) για την επίτευξη ενός στόχου για τον οποίο έχει κατασκευαστεί.

Η πολιτική (*policy*) προσδιορίζει τον τρόπο με τον οποίο ο πράκτορας ενεργεί σε μια δεδομένη χρονική στιγμή. Είναι μια απεικόνιση των καταστάσεων του περιβάλλοντος σε ενέργειες που μπορούν να επιλεχθούν όταν βρισκόμαστε σε αυτές τις καταστάσεις. Η πολιτική αποτελεί το κυριότερο στοιχείο ενός πράκτορα που χρησιμοποιεί την ενισχυτική μάθηση μιας και αυτή μπορεί να καθορίσει την συμπεριφορά του.

Η συνάρτηση ανταμοιβής (*reward function*) ορίζει τον στόχο σε ένα πρόβλημα ενισχυτικής μάθησης. Απεικονίζει κάθε ζεύγος κατάστασης-ενέργειας του περιβάλλοντος σε έναν αριθμό, την ανταμοιβή, η οποία εκφράζει το ποσο επιθυμητή είναι η συγκεκριμένη κατάσταση για τον πράκτορα. Ο στόχος του πράκτορα είναι να μεγιστοποιήσει μακροπρόθεσμα την συνολική ανταμοιβή. Η συνάρτηση ανταμοιβής ορίζει εν τέλει τη συμπεριφορά για έναν πράκτορα.

Η συνάρτηση αξίας (*value function*) είναι η συνολική ανταμοιβή που λαμβάνει ένας πράκτορας στο μέλλον, ξεκινώντας από την συγκεκριμένη κατάσταση. Ενώ η ανταμοιβή καθορίζει την άμεση πραγματική καταλληλότητα των καταστάσεων του περιβάλλοντος, η συνάρτηση αξίας δείχνει την μακροπρόθεσμη καταλληλότητα των καταστάσεων αφού λαμβάνει υπόψη τις καταστάσεις που είναι πιθανό να ακολουθήσουν και τις διαθέσιμες ανταμοιβές των καταστάσεων αυτών. Για παράδειγμα, μια κατάσταση μπορεί να αποφέρει μια μικρή ανταμοιβή αλλά ταυτόχρονα μπορεί να έχει μεγάλη αξία επειδή ακολουθείται συνεχώς από καταστάσεις που αποφέρουν μεγάλες ανταμοιβές. Αντίθετα με τις ανταμοιβές που δίνονται κατευθείαν από το περιβάλλον, οι αξίες των καταστάσεων πρέπει να εκτιμηθούν από τις ακολουθίες των παρατηρήσεων που κάνεις ένας πράκτορας σε όλη την διάρκεια της αλληλεπιδρασής του με το περιβάλλον.

Το μοντέλο (*model*) του περιβάλλοντος αποτελείται από μια συνάρτηση η οποία δέχεται ως όρισμα την τρέχουσα κατάσταση καθώς και μια ενέργεια κι επιστρέφει τη νέα κατάσταση στην οποία μεταβαίνει ο πράκτοραας και την ανταμοιβή. Ο πράκτορας συνήθως δεν έχει καμια γνώση του μοντέλου του περιβάλλοντος.




Σχήμα 2.2: Το πλαίσιο της ενισχυτικής μάθησης

## 2.2 Το πλαίσιο της Ενισχυτικής Μάθησης

Το πρόβλημα της ενισχυτικής μάθησης σχετίζεται με τον τρόπο που ένας πράκτορας μπορεί να μάθει μια συμπεριφορά μέσω της αλληλεπίδρασής του με το περιβάλλον για την επίτευξη ενός στόχου. Ο πράκτορας αλληλεπιδρά συνεχως με το περιβάλλον, επιλέγει ενέργειες και το περιβάλλον ανταποκρίνεται σε αυτές ανταμοίβοντας ή τιμωρόντας την τρέχουσα επιλογή του παρουσιάζοντάς του καινούργιες καταστάσεις.

Ο πράκτορας αλληλεπιδρά με το περιβάλλον κάθε διακριτή χρονική στιγμή, $t$ και λαμβάνει μια αναπαράσταση της κατάστασης του περιβάλλοντος, $s_t \in S$, όπου $S$ το σύνολο όλων των καταστάσεων. Με βάση την τρέχουσα κατάσταση επιλέγει μια ενέργεια $\alpha_t \in$ όπου $(s_t)$ είναι το σύνολο των διαθέσιμων ενεργειών στην συγκεκριμένη κατάσταση $s_t$. Την επόμενη χρονική στιγμή $(t+1)$, ο πράκτορας λαμβάνει μια ανταμοιβή $r_{t+1} \in R$ από το περιβάλλον ως συνέπεια της ενέργειάς του και μεταβαίνει σε μια νέα κατάσταση $s_{t+1}$. Στο Σχήμα 2.2 αναπαριστάται η αλληλεπίδραση πράκτορα και περιβάλλοντος.

Σε κάθε χρονική στιγμή ο πράκτορας απεικονίζει τις καταστάσεις σε πιθανότητες επιλογής όλων των δυνατών ενεργειών. Αυτή η απεικόνιση καλείται πολιτική (*policy*) του πράκτορα και συμβολίζεται ως $_t$, όπου $_t(s,a)$ είναι η πιθανότητα ο πράκτορας την χρονική στγμή $t$ να επιλέξει την ενέργεια $_t =$ αν βρίσκεται στην κατάσταση $s_t = s$. Οι μέθοδοι ενισχυτικής μάθησης προσδιορίζουν τον τρόπο με τον οποίο ο πράκτορας αλλάζει την πολιτική του ως αποτέλεσμα της εμπειρίας του. Στόχος

του είναι να μεγιστοποιήσει την συνολική ανταμοιβή που λαμβάνει μακροπρόθεσμα. Αυτό που προτείνει το πλαίσιο της ενισχυτικής μάθησης είναι ότι οποιοδήποτε πρόβλημα μάθησης συμπεριφοράς, οδηγούμενης από κάποιο συγκεκριμένο στόχο, μπορεί να αναχθεί σε τρία σήματα τα οποία ανταλλάσονται μεταξύ του πράκτορα και του περιβάλλοντος. Ένα σήμα για να αναπαρασταθούν οι επιλογές του πράκτορα (**ενέργειες**), ενα για να αναπαρασταθεί η πληροφορία στην οποία βασίζονται οι επιλογές των ενεργείών του (**καταστάσεις**) και τέλος ένα σήμα για να προσδιοριστεί ο στόχος του (**ανταμοιβή**).

## 2.3 Μοντελοποίηση Προβλημάτων Ενισχυτικής Μάθησης

Σε ένα πρόβλημα ενισχυτικής μάθησης οι ενέργειες που εκτελεί ο πράκτορας δεν καθορίζουν μόνο την άμεση ανταμοιβή που λαμβάνει από το περιβάλλον αλλά και την επόμενη κατάσταση στην οποία θα μεταβεί. Ενα τέτοιο περιβάλλον μπορεί να περιγραφεί ως ένα δίκτυο στο οποίο ο πράκτορας λαμβάνει υπόψη την επόμενη κατάσταση αλλά και την άμεση ανταμοιβή για να αποφασίσει ποια ενέργεια τελικά θα επιλέξει. Για τον λόγο αυτό το περιβάλλον μοντελοποιείται ως Μαρκοβιανές Διαδικασίες Απόφασης (*Markov Decision Processes - MDP*). Μια Μαρκοβιανή Διαδικασία Απόφασης περιγράφεται από μια πλειάδα $\{S, A, P, R\}$ όπου :

- $S$ αντιστοιχεί στο πεπερασμένο σύνολο όλων των δυνατών καταστάσεων,
- $A$ αντιστοιχεί στο πεπερασμένο σύνολο όλων των δυνατών ενεργειών,
- $P : S \times A \mapsto P(S)$ είναι η συνάρτηση μετάβασης όπου $P(S)$ είναι μια κατανομή πιθανοτήτων στο σύνολο των καταστάσεων $S$ η οποία δεδομένης μιας κατάστασης και μιας ενέργειας μας επιστρέφει τις πιθανότητες μετάβασης σε κάθε πιθανή επόμενη κατάσταση. Συμβολίζουμε ως $P^{\alpha}_{SS'}$ την πιθανότητα μετάβασης από την κατάσταση $s$ στην κατάσταση $s$' πραγματοποιώντας την ενέργεια $\alpha$,
- $R : S \times A \times S \mapsto \Re$ είναι η συνάρτηση ανταμοιβής, η οποία καθορίζει την επόμενη αναμενόμενη ανταμοιβή ως συνάρτηση της τρέχουσας κατάστασης και ενέργειας καθώς και της επόμενης κατάστασης. Συμβολίζουμε ως $R^{\alpha}_{SS'}$ την αναμενόμενη ανταμοιβή που θα πάρουμε αν στην κατάσταση $s$ επιλέξουμε την ενέργεια $\alpha$ και μεταβούμε στην κατάσταση $s'$.

Μια πολύ σημαντική υπόθεση πάνω στην οποία βασίζονται πολλά θεωρητικά αποτελέσματα είναι η ιδιότητα *Markov*. Αυτή η ιδιότητα αναφέρεται στο πως περιγράφεται η τρέχουσα κατάσταση. Ικανοποιείται εάν σε κάθε βήμα, περιλαμβάνεται όλη η πληροφορία που είναι απαραίτητη για τη διαδικασία λήψης απόφασης από τον πράκτορα. Με απλά λόγια, μας είναι αρκετή η γνώση της τρέχουσας κατάστασης και της επιλεχθείσας ενέργειας ώστε να καθοριστεί η επόμενη κατάσταση και ανταμοιβή. Η ιδιότητα *Markov* μπορεί να περιγραφεί με την ακόλουθη εξίσωση:

$$P\{s_{t+1}, r_{t+1} \mid s_t, t\} = P\{s_{t+1}, r_{t+1} \mid s_t, t, s_{t-1}, t-1, ..., s_0, \alpha_0\} \tag{2.1}$$

## 2.4 Συναρτήσεις Αξίας (Value Functions)

Στόχος του πράκτορα όπως προαναφέρθηκε είναι η μεγιστοποίηση της συνολικής ανταμοιβής που λαμβάνει μακροπρόθεσμα. Γενικά, ο πράκτορας θέλει να μεγιστοποιήσει την αναμενόμενη ανταμοιβή όπου η απολαβή $R_t$ ορίζεται ως μια συνάρτηση της ακολουθίας ανταμοιβών. Αν οι απολαβές που λαμβάνονται έπειτα από την χρονική στιγμή $t$ συμβολίζονται ως $r_{t+1}, r_{t+2}, r_{t+3}, ...$, ο πράκτορας θέλει να μεγιστοποιήσει το ακόλουθο άθροισμα ανταμοιβών:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + ... + r_T, \tag{2.2}$$

όπου  είναι το τελικό χρονικό βήμα. Αυτή η προσέγγιση έχει νόημα σε επεισοδιακές εφαρμογές όπου υπάρχει η έννοια του τελικού χρονικού βήματος. Επεισόδιο ορίζεται ως η ακολουθία των ενεργειών που εκτελούνται από εναν πράκτορα για να φθάσει από μια αρχική κατάσταση σε μια τελική. Κάθε επεισόδιο τερματίζει σε μια ειδική κατάσταση που ονομάζεται τερματική κατάσταση όπου όλες οι ενέργεις οδηγούν στην ίδια κατάσταση λαμβάνοντας μηδενική ανταμοιβή. Στη συνέχεια ο πράκτορας μεταβαίνει ξανά στην αρχική κατάσταση ή σε κάποια από τις αρχικές καταστάσεις με την ίδια πιθανότητα και ξεκινά ένα καινούργιο επεισόδιο. Σε προβλήματα όπου ο πράκτορας λαμβάνει ανταμοιβή μόνο όταν φθάνει στην τελική κατάσταση είναι συνηθισμένο να θεωρείται η τελική κατάσταση ως στόχος. Μερικές φορές όμως είναι αναγκαίο να διακρίνουμε το σύνολο των μη τερματικών καταστάσεων συμβολίζοντας το με $S^+$ ώστε να το διακρίνουμε από το σύνολο $S$ που είναι αυτό όλων των καταστάσεων.

Αντίθετα, σε πολλές περιπτώσεις η αλληλεπιδραση του πράκτορα με το περιβάλ-

λον δεν διακόπτεται σε διακεκριμένα επεισόδια αλλά συνεχίζεται επ' άπειρο χωρίς κάποιο περιορισμό. Αυτή την εργασία την ονομάζουμε συνεχόμενη. Ο ορισμός της (2.2) είναι προβληματικός για συνεχόμενες εργασίες διότι το τελικό χρονικό βήμα θα είναι $= \infty$ και η απολαβή μπορεί εύκολα να γίνει ίση με το άπειρο από μόνη της. Για το λόγο αυτό η (2.2) θα πρέπει να σταθμιστεί. Η επιπλέον ιδέα που εισαγουμε είναι αυτη της έκπτωσης. Σύμφωνα με αυτή τη προσέγγιση ο πράκτορας προσπαθεί να επιλέξει ενέργειες έτσι ώστε το άθροισμα των εκπτωθεισών ανταμοιβών που λαμβάνει να μεγιστοποιείται. Συγκεκριμένα επιλέγει την ενέργεια $\alpha_t$ έτσι ώστε να μεγιστοποιήσει την αναμενόμενη εκπτωθείσα απολαβή:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k r_{k+t+1}, \tag{2.3}$$

όπου $0 \leq \gamma \leq 1$ είναι ο ρυθμός έκπτωσης (*discount rate*).

Ο ρυθμός έκπτωσης καθορίζει την παρούσα αξία των μελλοντικών ανταμοιβών. Μια ανταμοιβή που λαμβάνεται $k$ χρονικά βήματα στο μέλλον αξίζει μόνο $\gamma^{k-1}$ φορές σε σχέση με την αξία που θα έχει αν ληφθεί άμεσα. Εάν $\gamma < 1$, το άπειρο άθροισμα έχει μια πεπερασμένη αξία όσο η ακολουθία των ανταμοιβών $r_k$ είναι οριοθετημένη. Εάν $\gamma = 0$, ο πράκτορας ενδιαφέρεται να μεγιστοποιήσει μόνο τις άμεσες ανταμοιβές και ο στόχος του είναι να μάθει πως θα επιλέξει την ενέργεια $\alpha_t$ ώστε να μεγιστοποιήσει μόνο το $r_{t+1}$. Καθώς το $\gamma$ προσεγγίζει το 1, οι μελλοντικές ανταμοιβές λαμβάνονται περισσότερο υπόψη με αποτέλεσμα ο πρακτορας να γίνεται περισσότερο προνοητικός. Η επιτυχία ενός πράκτορα αξιολογείται από το πόσο καλά μπορεί να μεγιστοποιεί μακροπρόθεσμα τη συνολική ανταμοιβή που λαμβάνει υπό μια πολιτική $\pi$. Η βέλτιστη πολιτική (*optimal policy*) $\pi^*$ , είναι μια πολιτική η οποία μεγιστοποιεί το αναμενόμενο μακροπρόθεσμα άθροισμα ανταμοιβών. Πολλές μέθοδοι ενισχυτικής μάθησης αντί να υπολογίζουν απευθείας την πολιτική, υπολογίζουν κάποια συνάρτηση αξίας και μετά την χρησιμοποιούν για να εξάγουν την πολιτική.

Η αξία μιας κατάστασης $s$ υπο μια πολιτική $\pi$ συμβολίζεται ως $V^\pi(s)$ και είναι η αναμενόμενη μέση απολαβή που λαμβάνει ο πράκτορας αν ξεκινήσει από την κατάσταση $s$ κι ακολουθήσει στη συνέχεια την πολιτική $\pi$. Για Μαρκοβιανές Διαδικασίες απόφασης ορίζουμε την συνάρτηση $V^\pi(s)$ ως :

$$V^\pi(s) = E_\pi\{R_t \mid s_t = s\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{k+t+1} \mid s_t = s \right\}, \tag{2.4}$$

με το $_\pi\{\}$ να υποδηλώνει την μέση τιμή, δεδομένου ότι ο πράκτορας ακολουθεί την πολιτική $\pi$. Η αξία κάθε τερματικής κατάστασης είναι πάντα μηδέν. Η $V^\pi(s)$ καλείται συνάρτηση αξίας κατάστασης (*state-value function*).

Ορίζουμε επίσης την $Q^\pi(s, \alpha)$ για τη λήψη μιας ενέργειας $\alpha$ στην κατάσταση $s$ υπό την πολιτική $\pi$ ως την αναμενόμενη απολαβή ξεκινώντας από την κατάσταση $s$, επιλέγοντας την ενέργεια $\alpha$ κι ακολουθώντας στη συνέχει την πολιτική $\pi$ :

$$Q^\pi(s, \alpha) = E_\pi\{R_t \mid s_t = s, \alpha_t = \alpha\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{k+t+1} \mid s_t = s, \alpha_t = \alpha\right\}, \qquad (2.5)$$

Τη συνάρτηση $Q^\pi(s, \alpha)$ την ονομάζουμε συναρτηση αξίας κατάστασης-ενέργειας (*action value function*).

Οι συναρτήσεις αξίας $V^\pi(s)$ και $Q^\pi(s, \alpha)$ μπορούν να υπολογιστούν εμπειρικά. Αν ο πράκτορας για παράδειγμα ακολουθεί μια πολιτικη $\pi$ και διατηρεί ένα μέσο όρο των πραγματικών απολαβών για κάθε κατάσταση που συναντά, τότε αυτός ο μέσος όρος θα συγκλίνει στην αξία της κατάστασης $V^\pi(s)$ καθώς ο αριθμός των επισκέψεων στην κατάσταση αυτή τείνει στο άπειρο. Επίσης αν διατηρούνται μέσοι όροι για κάθε ενέργεια που εκτελείται σε κάθε κατάσταση τότε αυτοί οι μέσοι όροι θα συγκλίνουν στις αξίες κατάστασης-ενέργειας $Q^\pi(s, \alpha)$. Ονομάζουμε λοιπόν τις παραπάνω μεθόδους ως μεθόδους *Monte Carlo*. Το πρόβλημα που συναντάμε σε αυτές τις μεθόδους είναι όταν το πλήθος καταστάσεων είναι μεγάλο. Τότε δεν είναι πρακτικό να κρατάμε μέσους όρους για κάθε κατάσταση ή ζεύγος κατάστασης-ενέργειας. Αντίθετα ο πράκτορας θα πρέπει να διατηρεί τις συναρτήσεις $V^\pi(s)$ και $Q^\pi(s, \alpha)$ ως παραμετροποιημένες συναρτήσεις προσαρμόζοντας τις παρεμέτρους έτσι ώστε να ταιριάζουν καλύτερα στις παρατηρούμενες απολαβές.

Μια ιδιότητα των συναρτήσεων αξίας είναι ότι ικανοποιούν συγκεκριμένες επαναληπτικές σχέσεις. Έτσι για μια οποιαδήποτε πολιτική $\pi$ και κατάσταση $s$ διατηρείται η ακόλουθη σχέση ανάμεσα στην αξία της κατάστασης $s$ και της αξίας της

πιθανής επόμενης κατάστασης:

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k}r_{k+t+1} \mid s_t = s,\right\} \\ &= E_{\pi}\left\{r_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^{k}r_{k+t+2} \mid s_t = s,\right\} \\ &= \sum_{\alpha}\pi(s,\alpha)\sum_{s'}P_{SS'}\left[R_{SS'} + \gamma E_{\pi}\left\{r_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^{k}r_{k+t+2} \mid s_{t+1} = s',\right\}\right] \\ &= \sum_{\alpha}\pi(s,\alpha)\sum_{s'}P_{SS'}\left[R_{SS'} + \gamma V^{\pi}(s')\right] \end{aligned} \tag{2.6}$$

Η εξίσωση 2.6 είναι η εξίσωση *Bellman* για την συνάρτηση αξίας κατάστασης $V^{\pi}(s)$ η οποία δηλώνει ότι η αξία της κατάστασης ισούται με την μειωμένη αξία της αναμενόμενης κατάστασης συν τη προσδοκόμενη ανταμοιβή.

Για πεπερασμένες Μαρκοβιανές Διαδικασίες Απόφασης μπορούμε να ορίσουμε τη βέλτιστη πολιτική με τον τρόπο που ακολουθεί. Μια πολιτική $\pi$ είναι καλύτερη ή ίση με μια πολιτική $\pi'$ εάν η μέση απολαβή της είναι καλύτερη ή ίση σε σχέση με αυτή της $\pi'$ για όλες τις καταστάσεις. Δηλαδή $\pi > \pi'$ αν και μόνο αν $V^{\pi}(s) \geq V^{\pi'}(s)$, $\forall s \in S$. Υπάρχει τουλάχιστον μια πολιτική η οποία είναι καλύτερη ή ίση σε σχέση με τις υπόλοιπες πολιτικές. Αυτή ονομάζεται βέλτιστη πολιτική. Μπορεί να υπάρχουν βέβαια και περισσότερες από μία βέλτιστες πολιτικές. Όλες τις συμβολίζουμε ως $\pi^*$. Η συνάρτηση αξίας κάταστασης ονομάζεται τώρα βελτιστη συνάρτηση αξίας κατάστασης $V^*$ και ορίζεται ως :

$$V^*(s) = \max_{\pi} V^{\pi}(s) \quad \text{για κάθε} \quad s \in S \tag{2.7}$$

Οι βέλτιστες πολιτικές μοιράζονται την ίδια συνάρτηση αξίας κατάστασης-ενέργειας $Q^*$ η οποία ορίζεται ως:

$$Q^*(s,\alpha) = \max_{\pi} Q^{\pi}(s,\alpha) \quad \text{για κάθε} \quad s \in \ \text{και} \ \alpha \in A \tag{2.8}$$

Για κάθε ζεύγος $(s,\alpha)$ η συνάρτηση αυτή δίνει την αναμενόμενη απολαβή που λαμβάνουμε αν στην κατάσταση $s$ επιλέξουμε την ενέργεια $\alpha$ και στη συνέχεια ακολουθήσουμε την βέλτιστη πολιτική. Έτσι μπορούμε να γράψουμε την $Q^*$ σε σχέση με την $V^*$ ως :

$$Q^*(s,\alpha) = E\{r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, \alpha_t = \alpha\} \tag{2.9}$$

Οι εξισώσεις $V^*$ και $Q^*$ καλούνται βέλτιστες εξισώσεις *Bellman*.
Η $V^*$ υπολογίζεται ως εξής :

$$\begin{aligned} V^*(s) &= \max_{\alpha \in A} Q^{\pi^*}(s,\alpha) \\ &= \max_{\alpha} E_{\pi^*}\{R_t \mid s_t = s, \alpha_t = \alpha\} \\ &= \max_{\alpha} E_{\pi^*}\left\{\sum_{k=0}^{\infty} \gamma^k r_{k+t+1} \mid s_t = s, \alpha_t = \alpha\right\} \\ &= \max_{\alpha} E_{\pi^*}\left\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{k+t+2} \mid s_t = s, \alpha_t = \alpha\right\} \\ &= \max_{\alpha} E\{r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, \alpha_t = \alpha\} \\ &= \max_{\alpha} \sum_{s'} P_{SS'}\Big[R_{SS'} + \gamma V^*(s')\Big] \end{aligned} \tag{2.10}$$

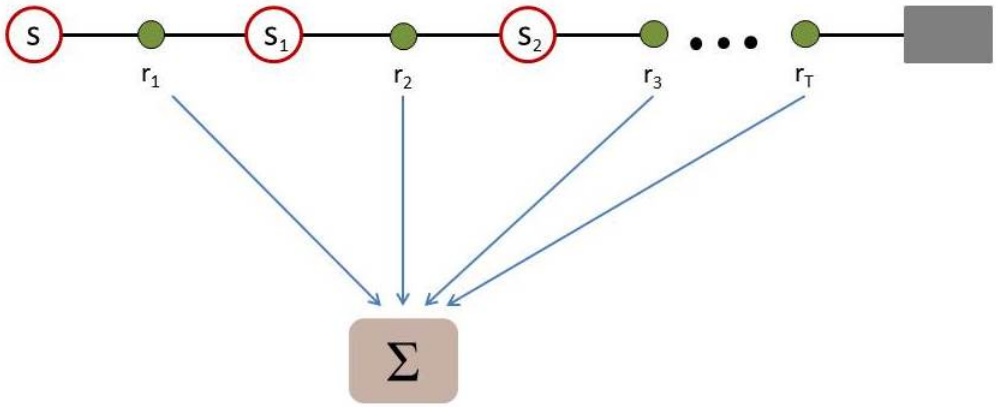
Η βέλτιστη εξίσωση *Bellman* για την $Q^*$ είναι η εξής :

$$\begin{aligned} Q^*(s,\alpha) &= E\{r_{t+1} + \gamma \max_{\alpha'} Q^*(s_{t+1},\alpha') \mid s_t = s, \alpha_t = \alpha\} \\ &= \sum_{s'} P_{SS'}\Big[R_{SS'} + \gamma \max_{\alpha'} Q^*(s',\alpha')\Big] \end{aligned} \tag{2.11}$$

Για πεπερασμένες *MDPs* η βέλτιστη συνάρτηση (2.10) έχει μια μοναδική λύση ανεξάρτητα από την πολιτική. Η βέλτιστη συνάρτηση Bellman, είναι ένα σύστημα εξισώσεων μία για κάθε κατάσταση. Έτσι, αν υπάρχουν καταστάσεις τότε υπάρχουν εξισώσεις με αγνώστους. Αν το μοντέλο του περιβάλλοντος είναι γνωστό τότε κάποιος μπορεί να λύσει αυτό το σύστημα εξισώσεων για την $V^*$ χρησιμοποιώντας μια μέθοδο επίλυσης μη γραμμικών εξισώσεων. Με ανάλογο τρόπο μπορεί να λυθεί και το σύστημα εξισώσεων για την $Q^*$. Εφόσον έχουμε ορίσει την $V^*$ είναι εύκολο να ορίσουμε και την βέλτιστη πολιτική. Για κάθε κατάσταση $s$ υπάρχουν μία ή περισσότερες ενέργειες που δίνουν την μέγιστη τιμή για την βέλτιση συνάρτηση *Bellman*. Εάν γνωρίζουμε την βέλτιστη συνάρτηση αξίας $V^*$ τότε οι ενέργειες που εμφανίζονται μετά από κάθε βήμα ως καλύτερες θα είναι οι βελτιστες. Η γνώση της $Q^*$ απλοποιεί περισσότερο την επιλογή των βέλτιστων ενεργειών. Για κάθε κατάσταση $s$ μπορεί να βρει οποιαδήποτε ενέργεια που μεγιστοποιεί την $Q^*(s,\alpha)$ χωρίς να ελέγχει το επόμενο βήμα. Με αυτό τον τρόπο, η βέλτιστη συνάρτηση αξίας κατάστασης-ενέργειας μας επιτρέπει να επιλέγουμε τις βέλτιστες ενέργειες δίχως να χρειάζεται γνώση για πιθανές διαδοχικές καταστάσεις και τις αξίες τους, δηλαδή γνώση του μοντέλου του περιβάλλοντος.

## 2.5 Μέθοδοι Monte Carlo

Οι μέθοδοι *Monte Carlo* δεν απαιτούν την γνώση του περιβάλλοντος. Βασίζονται μόνο στην συλλεγόμενη εμπειρία τους δηλαδή, στις ακολουθίες καταστάσεων, ενεργειών και ανταμοιβών που αποκτήθηκε από είτε από την *on-line* είτε από την *off-line* αλληλεπίδραση με το περιβάλλον. Η *on-line* μάθηση δεν απαιτεί καμία προηγούμενη γνώση του περιβάλλοντος και μπορεί να επιτύχει βέλτιση συμπεριφορά. Η *off-line* μάθηση δεν απαιτεί το πλήρες μοντέλο του περιβάλλοντος, αλλά αρκεί και μια προσέγγιση αυτού για να παραχθούν οι κατάλληλες πληροφορίες. Αυτές οι μέθοδοι είναι τρόποι επίλυσης προβλημάτων ενισχυτικής μάθησης και βασίζονται στους μέσους όρους των ανταμοιβών του πράκτορα. Για να εξασφαλίσουμε ότι οι ανταμοιβές είναι καλά ορισμένες ορίζουμε τις μεθόδους *Monte Carlo* μόνο για επεισοδιακές διεργασίες. Οι εκτιμήσεις των αξιών και οι πολιτικές μεταβάλλονται μόνο μετά την ολοκλήρωση ενός επεισοδίου. Στο Σχήμα 2.3 φαίνεται η απολαβή που λαμβάνει ο πράκτορας από τη στιγμή που θα συναντήσει την κατάσταση $s$ για πρώτη φορά μέχρι μια τερματική κατάσταση.




Σχήμα 2.3: Η απολαβή που λαμβάνει ο πράκτορας για ένα επεισόδιο

Για την εκτίμηση μιας πολιτικής οι μέθοδοι *Monte Carlo* χρησιμοποιούν την συνάρτηση αξίας κατάστασης, η οποία εκφράζει την αναμενόμενη απολαβή που θα λάβει ο πράκτορας ξεκινώντας απο αυτή την κατάσταση. Για τον υπολογισμό της υπολογίζουμε το μέσο όρο των απολαβών που παρατηρήθηκαν μετά από κάθε επίσκεψη στη συγκεκριμένη κατάσταση. Όσες περισσότερες είναι οι απολαβές που παρατηρούμε, τόσο περισσότερο ο μέσος όρος συγκλίνει στην αναμενόμενη αξία. Κάθε εμφάνιση της κατάστασης $s$ σε ένα επεισόδιο ονομάζεται επίσκεψη της $s$.

Όταν το μοντέλο του περιβάλλοντος δεν είναι διαθέσιμο χρησιμοποιούμε τις αξίας κατάστασης-ενέργειας αντί για τις αξίες κατάστασης. Γνωρίζοντας το μοντέλο οι αξίες κατάστασης είναι αρκετές για να προσδιορίσουμε μια πολιτική. Θα πρέπει όμως να εκτιμήσουμε την αξία κάθε κατάστασης για να ορίσουμε μια πολιτική. Συνεπώς ένας από τους κυριότερους στόχους για τις μεθόδους *Monte Carlo* είναι η εκτίμηση της $Q^*$. Το πρόβλημα της εύρεσης της πολιτικής για τις αξίες κατάστασης-ενέργειας είναι η εκτίμηση της $Q^\pi(s, \alpha)$ δηλαδή της απολαβής που αναμένουμε ξεκινώντας από την κατάσταση $s$, επιλέγοντας την ενέργεια $\alpha$ και στη συνέχεια ακολουθώντας την πολιτική $\pi$. Το πρόβλημα που προκύπτει είναι ότι ο πράκτορας μπορεί να μην επισκευφτεί ποτέ ορισμένα ζεύγη κατάστασης-ενέργειας. Όταν η πολιτική $\pi$ είναι ντετερμινιστική, ακολουθώντας την θα παρατηρήσει κάποιος απολαβές μόνο για μια ενέργεια για κάθε κατάσταση. Για να υπολογίστει ο μέσος όρος όταν δεν υπάρχουν απολαβές, οι εκτιμήσεις *Monte Carlo* για τις άλλες ενέργειες δε θα βελτιωθούν ποτέ με την εμπειρία. Αυτό είναι ενα σημαντικό πρόβλημα μιας και ο σκοπός της μάθησης των αξιών κατάστασης-ενέργειας είναι να βοηθήσει στην επιλογή ανάμεσα στις ενέργειες που είναι διαθέσιμες σε κάθε κατάσταση. Ίναι αναγκαίο δηλαδή να υπολογίσουμε τις αξίες όλων των ενεργειών για κάθε κατάσταση.

Για να λειτουργήσει η εκτίμηση της πολιτικής για τις αξίες κατάστασης-ενέργειας θα πρέπει να υποθέσουμε πλήρη εξερεύνηση. Ενας τρόπος για να πραγματοποιηθεί αυτό είναι υποθέτοντας ότι το πρώτο βήμα σε κάθε επεισόδιο ξεκινά από ένα ζεύγος κατάστασης-ενέργειας και κάθε ζεύγος αυτής της μορφής έχει μη μηδενική πιθανότητα για να επιλεχθεί όταν ξεκινήσει το επεισόδιο. Αυτός ο τρόπος μας εγγυάται ότι όλα τα ζεύγη κατάστασης-ενέργειας είναι επισκέψιμα άπειρες φορές καθώς ο αριθμός των επεισοδίων τείνει στο άπειρο. Αυτό ονομάζεται εξερεύνηση αφετηρίας. Η υπόθεση της εξερεύνησης αφετηρίας είναι μερικές φορές χρήσιμη αλλα φυσικά δεν μπορούμε αυτό να το γενικεύσουμε ιδιαίτερα όταν η μάθηση γίνεται με πραγματική αλληλεπίδραση με ένα περιβάλλον. Σε αυτή την περίπτωση, αυτό το ζεύγος κατάστασης-ενέργειας δεν είναι ιδιαίτερα χρήσιμο.

Η βελτίωση της πολιτικής επιτυχγάνεται κάνοντας την πολιτική άπληστη σε σχέση με την τρέχουσα συνάρτηση αξίας. Στη συγκεκριμένη περίπτωση η συνάρτηση αξίας κατάστασης-ενέργειας είναι διαθέσιμη με αποτέλεσμα το μοντέλο για την δημιουργία της άπληστης πολιτικής να μη χρειάζεται. Για κάθε συνάρτηση αξίας κατάστασης-ενέργειας $Q$, η αντίστοιχη άπληστη πολιτική είναι αυτή που για κάθε

κάτασταση $s \in S$ ντετερμινιστικά επιλέγει αυτή την ενέργεια με την μεγαλύτερη αξία $Q$:

$$\pi(s) = arg\max_{\alpha} Q(s, \alpha) \tag{2.12}$$

Για την εκτίμηση της πολιτικής *Monte Carlo* είναι φυσιολογική η εναλλαγή ανάμεσα στην εκτίμηση και τη βελτίωση ανά επεισόδιο. Μετά το τέλος του κάθε επεισοδίου οι παρατηρούμενες απολαβές χρησιμοποιούνται για την εκτίμηση της πολιτικής, ενώ αμέσως μετά η πολιτική βελτιώνεται σε όλες τις καταστάσεις που έχουν επισκεφτεί κατά τη διάρκεια του επεισοδίου.

## 2.6 Μάθηση Χρονικών Διαφορών

Η μάθηση Χρονικών Διαφορών (*Temporal-Difference Learning - TD*) αποτελεί έναν συνδιασμό των μεθόδων *Monte Carlo* και Δυναμικού Προγραμματισμού. Όπως και στις μεθόδους *Monte Carlo*, οι μέθοδοι Χρονικών Διαφορών δεν χρειάζονται την γνώση του μοντέλου του περιβάλλοντος. Κάνουν ενημέρωση των εκτιμήσεων βασιζόμενοι σε ήδη γνωστές εκτιμήσεις δίχως να περιμένουν την τελική απολαβή. Τόσο οι μέθοδοι Χρονικών Διαφορών όσο και οι μέθοδοι *Monte Carlo* χρησιμοποιούν την εμπειρία για την επίλυση του προβλήματος της πρόβλεψης. Έχοντας κάποια εμπειρία ακολουθόντας μια πολιτική $\pi$ και οι δύο προαναφερθείσες μέθοδοι ενημερώνουν τις εκτιμήσεις $V$ της $V^{\pi}$. Αν επισκεφθούμε μια μη τερματική κατάσταση $s_t$ τη χρονική στιγμή $t$, τότε οι δύο μέθοδοι ενημερώνουν την εκτίμηση $V(s_t)$ βασιζόμενες στο τι θα συμβεί μετά την επίσκεψη. Οι μέθοδοι *Monte Carlo* περιμένουν μέχρι η απολαβή που ακολουθεί την επίσκεψη να γίνει γνωστή κι έπειτα τη χρησιμοποιούν ως στόχο για την $V(s_t)$. Αντίθετα οι μέθοδοι Χρονικών Διαφορών χρειάζονται να περιμένουν μόνο μέχρι το επόμενο βήμα. Η πιο απλή μέθοδος χρονικών διαφορών είναι η *TD(0)* της οποίας ο κανόνας ενημέρωσης είναι ο εξής:

$$V(s_t) \leftarrow V(s_t) + \alpha \underbrace{[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]}_{\text{TD σφάλμα}}. \tag{2.13}$$

Το $r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ ονομάζεται σφάλμα χρονικών διαφορών, το $\alpha$ αποτελεί τον ρυθμό μάθησης (*learning rate*),$0 < \alpha < 1$ και το $\gamma$, $0 < \gamma < 1$ το ρυθμό έκπτωσης (*discount rate*). Αν τη χρονική στιγμή $t$ επισκεφθούμε την κατάσταση $s_t$, τότε η εκτιμώμενη αξία της ενημερώνεται έτσι ώστε να είναι πιο κοντά στην $r_{t+1} + \gamma V(s_{t+1})$, όπου $r_{t+1}$ είναι η άμεση ανταμοιβή που λαμβάνεται και $V(s_{t+1})$ η εκτιμώμενη αξία της

επόμενης κατάστασης. Η κύρια ιδέα είναι ότι η $r_{t+1} + \gamma V(s_{t+1})$ είναι ένα δείγμα της $V(s_t)$ και είναι περισσότερο πιθανό να είναι σωστή διότι ενσωματώνει την πραγματική άμεση ανταμοιβή $r_{t+1}$. Πιο συγκεκριμένα η ποσότητα $r_{t+1} + \gamma V(s_{t+1})$ είναι αυτή ως προς την οποία θέλουμε να μετατοπίσουμε την αξίας της κατάστασης $s_t$. Με την κατάλληλη προσαρμογή του ρυθμού μάθησης είναι σίγουρο ότι ο *TD(0)*(Αλγόριθμος 1) θα συγκλίνει στην βέλτιστη συνάρτηση αξίας.

Ένα πλεονέκτημα των μεθόδων χρονικών διαφορών σε σχέση με τις μεθόδους *Monte Carlo* είναι ότι υλοποιούνται επαυξητικά μιας και θα πρέπει να περάσει μόνο ένα χρονικό βήμα για να γίνει η ενημέρωση σε αντίθεση με τις μεθόδους *Monte Carlo* που θα πρέπει να περιμένουμε μέχρι το τέλος του επεισοδίου, αφού τότε γίνεται γνωστή η απολαβή.

**Αλγόριθμος 2.1** Ο αλγόριθμος TD(0)

Αρχικοποίηση τυχαία την $V(s)$
$\pi$ η πολιτική προς εκτίμηση
**repeat** (για κάθε επεισόδιο)
  Αρχικοποίηση της s
  **repeat** (για κάθε βήμα του επεισοδίου)
    $\alpha \leftarrow$ ενέργεια που επιλέγεται από την πολιτική $\pi$ για την κατάσταση $s$
    Εκτέλεση της ενέργειας $\alpha$
    Λήψη της ανταμοιβή $r$ από το περιβάλλον
    Μετάβαση στην νέα κατάσταση $s'$
    $V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)]$
    $s \leftarrow s'$
  **until** $s$ είναι μια τερματική κατάσταση

### 2.6.1 Ο αλγόριθμος Q-Learning

Ο αλγόριθμος *Q-learning* (Watkins, 1989; Watkins & Dayan, 1992) είναι ένας από τους πιο γνωστους αλγορίθμους της ενισχυτικής μάθησης που χρησιμοποιείται σε προβλήματα ελέγχου. Είναι *off-line* αλγόριθμος δηλαδή η πολιτική που χρησιμοποιείται για την λήψη αποφάσεων δεν χρειάζεται να είναι ίδια με αυτή που αξιολογείται και βελτιώνεται. Καθώς είναι ένας αλγόριθμος που δεν χρειάζεται να γνωρίζει το μοντέλο του περιβάλλοντος η συνάρτηση αξίας ενέργειας-κατάστασης εκτιμάται

αντί για την συνάρτηση αξίας κατάστασης ώστε να φθάσει στην ανακάλυψη της βέλτιστης πολιτικής. Οι τιμές της συνάρτησης $Q$ αρχικοποιούνται τυχαία. Η διαδικασία μάθησης πραγματοποιειται σε πραγματικό χρόνο. Σε κάθε χρονικό βήμα $t$ ο πράκτορας παρατηρεί μια κατάσταση $s_t$ και λαμβάνει μια ενέργεια $\alpha_t$ σύμφωνα με την πολιτική. Ως αποτέλεσμα αυτής της ενέργειας που έλαβε είναι η μετάβαση του σε μια ένα κατάσταση $s_{t+1}$ λαμβάνοντας μια άμεση ανταμοιβή από το περιβάλλον. Ο κανόνας ενημέρωσης της συνάρτησης αξίας κατάστασης-ενέργειας δίνεται παρακάτω:

$$Q^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) + \eta(r_t + \gamma \max_{\alpha \in A} Q^{\pi}(s_{t+1}, a) - Q^{\pi}(s_t, a_t)) \tag{2.14}$$

όπου $\eta \in (0, 1]$ και είναι η παράμετρος του ρυθμού μάθησης. Ο Αλγόριθμος 2 παρουσιάζει τη μέθοδο σε μορφή ψευδοκώδικα.

**Αλγόριθμος 2.2** Q-Learning

Αρχικοποίηση τυχαία της $Q(s, a)$
**repeat** (για κάθε επεισόδιο)
  Αρχικοποίηση της κατάστασης $s$
  **repeat** (για κάθε βήμα του επεισοδίου)
    Επιλογή της ενέργειας $\alpha$ στην $s$ με χρήση της πολιτική $\pi$ από την $Q$
    Εκτέλεση της ενέργεια $\alpha$
    Λήψη της ανταμοιβής $r$ από το περιβάλλον
    Μετάβαση στην νέα κατάσταση $s'$
    $Q(s, \alpha) \leftarrow Q(s, \alpha) + \eta[r + \gamma Q(s', \alpha') - Q(s, \alpha)]$
    $s \leftarrow s'$
  **until** $s$ είναι μια τερματική κατάσταση
**until** (σύγκλιση)

Στη συγκεκριμένη περίπτωση η συνάρτηση αξίας κατάστασης-ενέργειας $Q$ προσεγγίζει άμεσα τη βέλτιστη συνάρτηση αξίας κατάστασης-ενέργειας $Q^*$, ανεξάρτητα από την πολιτική που ακολουθείται. Αν κάθε ενέργεια εκτελείται σε κάθε κατάσταση άπειρες φορές και ο ρυθμός μάθησης φθίνει κατάλληλα τότε οι αξίες της $Q$ θα συγκλίνουν με πιθανότητα 1 στην $Q^*$.

## 2.7 Προσέγγιση Συνάρτηση Αξίας (Value Function Approximation)

Η πλειοψηφία των μεθόδων ενισχυτικής μάθησης βασίζονται στην εκτίμηση μιας συνάρτησης αξίας κατάστασης ή κατάστασης-ενέργειας. Όταν ο χώρος καταστάσεων είναι διακριτός η συνάρτηση αξίας κατάστασης μπορεί να αναπαρασταθεί με την χρήση ενός πίνακα που περιέχει μια εγγραφή για κάθε διακριτή κατάσταση. Στην περίπτωση όμως που έχουμε να επιλύσουμε προβλήματα με μεγάλο χώρο καταστάσεων ή άπειρο η συνάρτηση αξίας δεν μπορεί να αναπαρασταθεί χρησιμοποιώντας μια δομή πίνακα. Το πρόβλημα δεν είναι μόνο η μνήμη που απαιτείται για την αποθήκευση μεγάλων πινάκων αλλα και ο χρόνος που τα δεδομένα χρειάζονται για να γεμίσουν τους πίνακες με τις σωστές τιμές. Για το λογο αυτό χρησιμοποιείται ένα σχήμα προσέγγισης της συνάρτησης. Η προσέγγιση συνάρτησης καθιστά πρακτική την αναπαράσταση συναρτήσεων αξίας για μεγάλους χώρους καταστάσεων. Μια από τις σημαντικότερες κατηγορίες μεθόδων προσέγγισης συνάρτησης είναι οι γραμμικές μέθοδοι. Σε αυτές τις μεθόδους τη χρονική στιγμή $t$ η συνάρτηση προς προσέγγιση, $V_t$, είναι γραμμική ως προς το διάνυσμα παραμέτρων $w_t$. Το διάνυσμα παραμέτρων ειναι ένα διάνυσμα στήλης $w^t = [{w_1}^t, {w_2}^t, \cdots, {w_k}^t]^\top$. Σε κάθε κατάσταση $s$ αντιστοιχεί ένα διάνυσμα χαρακτηριστικών $\phi(s) = [\phi_1(s), \phi_2(s), \cdots, \phi_k(s)]^\top$ με ίδιο αριθμό παραμέτρων με το $w$. Επομένως η συνάρτηση αξίας κατάστασης αναπαριστάται ως εξης:

$$V(s) = \sum_{i=1}^{k} w_i \phi_i(s) = \phi(s)^\top w, \tag{2.15}$$

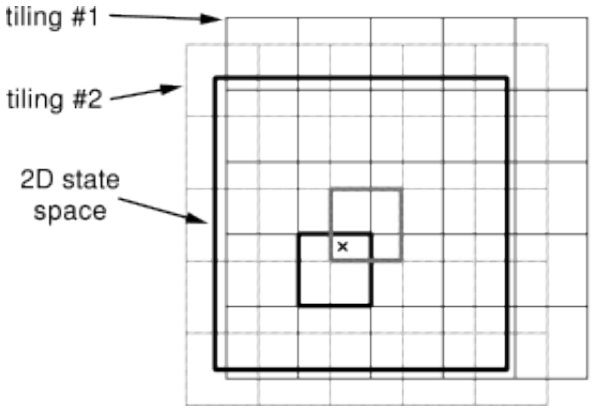
Η κατάλληλη εισαγωγή χαρακτηριστικών είναι ένας τρόπος για την εισαγωγή προγενέστερης αποκτηθείσας γνώσης στα συστήματα ενισχυτικής μάθησης. Τα χαρακτηριστιά θα πρέπει να αντιστοιχούν στα φυσικά χαρακτηριστικά της εργασίας. Στη συνέχεια εξετάζουμε τρόπους για την κατασκευή του χώρου των χαρακτηριστικών.

### 2.7.1 Κατασκευή χώρου χαρακτηριστικών

**Κωδικοποίηση Πλακιδίου (*Tile Coding*)**

Η ιδέα πίσω από την κωδικοποίηση πλακιδίου περιλαμβάνει την χρήση πολλαπλών μη επικαλυπτόμενων διαμερίσεων του χώρου καταστάσεων γνωστά ως πλακίδια. Η μέθοδος αυτή δίνει τη δυνατότητα της διακριτοποίησης ενός συνεχούς χώρου καταστάσεων σε πλακίδια. Αυξάνοντας το πλάτος των πλακιδίων έχει ως αποτέλεσμα

καλύτερη ικανότητα γενίκευσης, ενώ ο αριθμός των επικαλυπτόμενων διαμερίσεων επηρεάζει την ακρίβεια της επιθυμητής συνάρτησης. Στο Σχήμα 2.4 παρουσιάζεται ένα παράδειγμα της κωδικοποίησης πλακιδίου για δύο μεταβλητές. Κάθε στοιχείο




Σχήμα 2.4: Κωδικοποίηση Πλακιδίου για δύο μεταβλητές

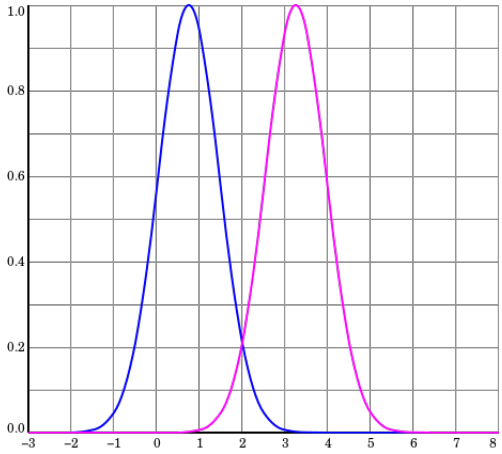
ενός πλακιδίου είναι ένα δυαδικό χαρακτηριστικό και ενεργοποιείται εάν και μόνο εάν η σχετική κατάσταση πέφτει μέσα στην περιοχή που οριοθετείται από το πλακίδιο. Όλα τα πλακίδια δεν κατανέμονται με τον ίδιο τρόπο αλλά με μια μικρή απόκλιση το ένα από το άλλο. Αυτό καθιστά αυτή την κωδικοποίηση μια αποτελεσματική προσεγγιστική μέθοδο.

**Ακτινικές Συναρτήσεις Βάσης (Radial Basis Functions - RBF)**

Οι ακτινικές συναρτήσεις βάσης, είναι ο πιο συνηθισμένος τρόπος κατασκευή του χώρου των χαρακτηριστικών για την προσέγγιση συναρτήσεων. Κάθε χαρακτηριστικό λαμβάνει συνεχείς τιμές στο διάστημα [0,1]. Με αυτόν τον τρόπο, αντικατοπτρίζει τους διαφορετικούς βαθμούς στους οποίους το χαρακτηριστικό συμμετέχει σε κάθε κατάσταση. Ένα τυπικό χαρακτηριστικό, $\phi_i$, ενός *RBF* υπολογίζεται σύμφωνα με τη συνάρτηση *Gauss*:

$$\phi_i(s) = \exp\left\{ -\frac{(s - c_i)^2}{2\sigma_i^2} \right\}, \tag{2.16}$$

όπου $c_i$ είναι το κέντρο κάθε συνάρτησης και το $\sigma^2$ είναι η διακύμανση που ορίζει το πλάτος της. Το Σχήμα 2.5 δείχνει ένα παράδειγμα με δύο ακτινικές συναρτήσεις βάσης με διαφορετικά κέντρα. Η επιλογή της τιμής του σ αλλά και το πλήθος ($k$) των συναρτήσεων *RBF* που θα χρησιμοποιούν αποτελούν κρίσιμες παραμέτρους. Ένας μικρός σε πλήθος χώρος χαρακτηριστικών μπορεί να μην επαρκεί για την σωστή προσέγγιση της συνάρτησης και να οδηγήσει σε *underfitting*. Αντίθετα, πολλά σε πλήθος χαρακτηριστικά μπορεί να προκαλέσουν *overfitting*. Ετσι το πλήθος ($k$) των χαρακτηριστικών πρέπει να επιλεγεί προσεκτικά.




Σχήμα 2.5: Συναρτήσεις ακτινικής βάσης στη μια διάσταση

## 2.8 Exploration *vs* Exploitation

Μια από τις προκλήσεις που προκύπτουν στην ενισχυτική μάθηση σε αντίθεση με άλλα είδη μάθησης είναι το δίλλημα της εξισορρόπησης (*trade-off*), μεταξύ εξερεύνησης (*exploration*) και αξιοποίησης (*exploitation*). Για την απόκτηση μεγαλύτερων ανταμοιβών, ο πράκτορας θα πρέπει να επιλέγει ενέργειες που επιλέχθηκαν στο παρελθόν και αποδείχθηκαν αποτελεσματικές στην παραγωγή ανταμοιβών. Για να ανακαλύψουμε τις ενέργειες αυτές θα πρέπει να δοκιμάσουμε ενέργειες που δεν έχουν ακόμη επιλεχθεί. Ο πράκτορας θα πρέπει να εκμεταλλευτεί την ήδη αποκτηθείσα γνώση του για να πάρει καλές ανταμοιβές και επίσης θα πρέπει να εξερευνεί ώστε να κάνει καλύτερες επιλογές ενεργειών στο μέλλον. Το θέμα που προκύπτει είναι με ποιό τρόπο θα πετύχουμε εξισορρόπηση μεταξύ της εξερεύνησης και της αξιοποίησης, δηλαδή κατά πόσο ο πράκτορας πρέπει να επιλέγει τυχαίες ενέργειες έτσι ώστε να μπορέσει να επισκεφθεί καινούργιες καταστάσεις ή να αξιοποιήσει την ήδη υπάρχουσα γνώση του ώστε να μεγιστοποιήσει την απολαβή του. Η εξερεύνηση είναι πιθανό να οδηγήσει τον πράκτορα σε καταστάσεις που θα του αποφέρουν ακόμα μεγαλύτερες ανταμοιβές σε σχέση με την αξιοποίηση. Όμως ο πράκτορας δεν μπορεί να κάνει εξερεύνηση επ άπειρον μιας και πρέπει να αξιοποιήσει την γνώση που έχει αποκτήσει ώστε να μεγιστοποιήσει τις απολαβές του.

### 2.8.1 $\epsilon$-greedy επιλογή ενέργειας

Μια απλή μέθοδος για την εξισορρόπηση μεταξύ εξερεύνησης κι αξιοποίησης είναι η $\epsilon$-greedy επιλογή ενέργειας. Χρησιμοποιώντας αυτή την μέθοδο μπορούμε να

επιλέξουμε μια τυχαία ενέργεια με πιθανότητα $\epsilon \in [0,1]$ ενώ με πιθανότητα $1-\epsilon$ να επιλέξουμε την ενέργεια εκείνη με την μεγαλύτερη ανταμοιβή σε μια δοθείσα κατάσταση. Το πλεονέκτημα της μεθόδου είναι πως κάθε ενέργεια θα επιλεχθεί τουλάχιστον μία φορά.

**Αλγόριθμος 2.3** $\epsilon$-greedy

**if** τυχαίος αριθμός στο (0,1) $\leq \epsilon$
  Επιλογή τυχαίας ενέργειας
**else**
  Επιλογή της καλύτερης ενέργειας
**end**

## 2.9 Η μέθοδος Least Squares Policy Iteration (LSPI)

Ο αλγόριθμος *LSPI* (Lagoudakis & Parr, 2003) είναι μια επαναληπτική διαδικάσία για την εκτίμηση της πολιτικής. Βασίζεται στη μέθοδο της εκτίμησης παραμέτρων των ελάχιστων τετραγώνων. Η συνάρτηση αξίας κατάστασης-ενέργειας για όλα τα ζεύγη (κατάστασης,ενέργειας) υπολογίζεται σύμφωνα με την εξίσωση *Bellman* και γράφεται ως :

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \gamma \sum_{s'} P^{\alpha}_{ss'} Q^{\pi}(s', \pi(s')) \tag{2.17}$$

Αλγεβρικά η εξίσωση 2.17 γράφεται

$$Q^{\pi} = R + \gamma P^{\pi} Q^{\pi} \tag{2.18}$$

όπου $Q^{\pi}$ και $R$ είναι διανύσματα μεγέθους $|S||A|$ και $P^{\pi}$ είναι ο στοχαστικός πίνακας μεταβάσεων με μέγεθος ($|S||A| \times |S||A|$ Όπως αναφέραμε και σε προηγούμενη ενότητα όταν ο χώρος καταστάσεων είναι πολύ μεγάλος ή άπειρος χρησιμοποιείται μια προσέγγιση της συνάρτησης κατάστασης ή κατάστασης-ενέργειας. Μια μέθοδος που χρησιμοποιείται συχνά είναι η χρήση του γραμμικού μοντέλου στο οποίο η συνάρτηση αξίας κατάστασης-ενέργειας εκτιμάτε ως γραμμικός συνδυασμός των $k$ παραμέτρων $w$ και των $k$ χαρακτηριστικών $\phi$.

$$Q^{\pi}(s,\alpha,w) = \sum_{i=1}^{k} \phi_i(s,\alpha) w_i = \phi(s,\alpha)^T w. \tag{2.19}$$

Σε πολλές εφαρμογές το μοντέλο$(R, P^{\pi})$ δεν είναι διαθέσιμο και η συνάρτηση αξίας για να εκτιμηθεί χρειάζεται ένα σύνολο από δείγματα δεδομένων. Αυτά τα δείγματα είναι πλειάδες της μορφής $D = \{s_i, a_i, r_i, s_i' | i = 1, \ldots, n\}$ όπου $s_i$ είναι μια κατάσταση, $_i$ μια ενέργεια που επιλέχθηκε με αποτέλεσμα μια μετάβαση σε μια νέα κατάσταση $s_i{}'$ και $r_i$ μια ανταμοιβή από το περιβάλλον. Αυτά τα δείγματα μπορεί να ληφθούν είτε απο μια πραγματική επεισοδιακή διαδικασία είτε με τυχαίο τρόπο. Οι πίνακες $\Phi, P\Phi = '$ και $R$ που προκύπτουν είναι:

$$\Phi = \begin{pmatrix} \phi(s_1, \alpha_1)^\top \\ \vdots \\ \phi(s, \alpha)^\top \\ \vdots \\ \phi(s_n, \alpha_n)^\top \end{pmatrix} \qquad \Phi' = \begin{pmatrix} \phi(s_1{}', \pi(s_1'))^\top \\ \vdots \\ \phi(s', \pi(s'))^\top \\ \vdots \\ \phi(s_n{}', \pi(s_n'))^\top \end{pmatrix} \qquad R = \begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix}$$

Συνδιάζοντας τις εξισώσεις 2.18 και 2.19 έχουμε:

$$\Phi w = R + \gamma P^{\pi} \Phi w \tag{2.20}$$

όπου Φ πίνακας μεγέθους $|S||A| \times k$
Ορίζεται έτσι το πρόβλημα εκτίμησης παραμέτρων ελαχίστων τετραγώνων:

$$\min_w \frac{1}{2} ||\Phi w - (R + \gamma \Phi' w)||^2 \tag{2.21}$$

Η λύση αυτού του προβλήματος προκύπτει από τον εκτιμητή ελαχίστων τετραγώνων θέτοντας την παράγωγο ως προς $w$ ίση με το μηδεν. Οπότε προκύπτει:

$$\widehat{w} = A^{-1} b \tag{2.22}$$

όπου

$$A = \Phi(\Phi - \gamma \Phi')^\top \tag{2.23}$$

$$b = \Phi R \tag{2.24}$$

με $A$ έναν $(k \times k)$ πίνακα και b ένα διάνυσμα $(k \times 1)$. Τα δείγματα $D$ που συλλέχθηκαν χρησιμοποιούνται για την εκτίμηση των παραμέτρων $w$ και για την εύρεση μιας πολιτικής. Έπειτα με βάση την τρέχουσα πολιτική δημιουργεί ένα νέο σύνολο δειγμάτων $D'$ και χρησιμοποιώντας το εκτιμά εκ νέου τις παραμέτρους $w$. Αυτή η διαδικασία επαναλαμβάνεται μέχρις ότου να ικανοποιείται ένα κριτήριο σύγκλισης.

Ο αλγόριθμος που παρουσιάστηκε παραπάνω είναι ένας *off-line* αλγόριθμος μιας και δέχεται ως είσοδο ένα σύνολο απο παραδείγματα έτσι ώστε να ανακαλύψει την βελτιστη πολιτική. Υπάρχει όμως και η *on-line* προσεγγίση του αλγόριθμου. Η κυριότερη διαφορά της *online* έκδοσης του *LSPI* με αυτή της *offline* είναι ότι δεν χρειάζεται ως είσοδο έναν αριθμό δειγμάτων καθώς αυτά εμφανίζονται ακολουθιακά κι ότι η πολιτική βελτιώνεται σε κάθε βήμα που εκτελεί ο πράκτορας ή έπειτα από έναν αριθμό βημάτων. Κατά τη διαδικασία εκτίμησης της πολιτικής ο παραπάνω πίνακας και το διάνυσμα $b$ μπορούν να γραφούν και με την ακόλουθη μορφή:

$$A = \sum_{i=1}^{n} \phi(s_i, \alpha_i)(\phi(s_i, \alpha_i) - \gamma\phi(s_i', \pi(s_i')))^\top \tag{2.25}$$

$$b = \sum_{i=1}^{n} \phi(s_i, \alpha_i) r_i \tag{2.26}$$

Αξιοποιώντας τις σχέσεις (2.26) και (2.27) καταλήγουμε στο συμπέρασμα ότι ο και $b$ μπορούν να δημιουργούνται αυξητικά σε κάθε βήμα του αλγορίθμου για την εκτίμηση των παραμέτρων $w$. Υποθέτωντας ότι αρχικοποιουνται και οι δύο με την τιμή μηδέν, για ένα νέο δείγμα τη χρονική στιγμή $t$ $(s_t, \alpha_t, r_t, s_t')$ υπολογίζονται ως :

$$A = A + \phi(s_t, \alpha_t)(\phi(s_t, \alpha_t) - \gamma\phi(s_t', \pi(s_t')))^\top \tag{2.27}$$

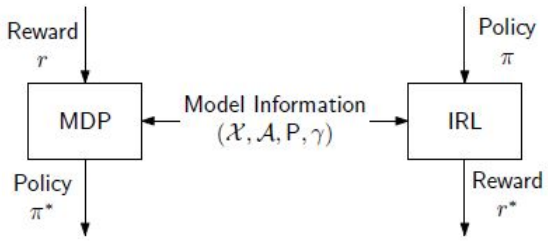
$$b = b + \phi(s_t, \alpha_t) r_t \tag{2.28}$$

Αυτό επιτρέπει την διαχείριση απεριοριστο αριθμό δειγμάτων και καθιστά εφικτή την *online* μάθηση.

## 2.10 Αντίστροφη Ενισχυτική Μάθηση

Η συνάρτηση ανταμοιβής αποτελεί σπουδαίο κομμάτι της ενισχυτικής μάθησης μιας και ορίζει τη συμπεριφορά του πράκτορας. Οι απολαβές πρέπει να είναι ορισμένες με τέτοιο τρόπο ώστε να κατευθύνουν τον μαθητευόμενο πράκτορα στην επίτευξη του στόχου του κι ο καθορισμός της αποτελεί μια δύσκολη διαδικασία. Το πρόβλημα του ορισμού της συνάρτησης ανταμοιβής καλείται να επιλύσει η αντίστροφη ενισχυτικής μάθηση (*Inverse Reinforcement Learning - IRL*). Οι μέθοδοι που υπάγονται σε αυτή την κατηγορία επικεντρώνονται στην εύρεση της συνάρτησης ανταμοιβής (*reward function*). Στο σημείο αυτό θα ορίσουμε το πρόβλημα της αντίστροφης ενισχυτικής μάθησης. Όπως είδαμε παραπάνω μια Μαρκοβιανή Διαδικασία Απόφασης

αντιπροσωπεύει ένα πρόβλημα απόφασης στο οποίο η διεργασία που καλείται να ολοκληρωθεί εκπροσωπείται από τη συνάρτηση ανταμοιβής $r$. Η βέλτιστη λύση σε μια τέτοια διεργασία καταλήγει σε μια πολιτική $\pi^*$. Η αντίστροφη ενισχυτική μά-




Σχήμα 2.6: Το πλαίσιο της αντίστροφης ενισχυτικής μάθησης

θηση πραγματεύεται με το αντίστροφο πρόβλημα από μια μαρκοβιανή διαδικασία απόφασης όπως παρουσιάζεται στο Σχήμα 2.6. Λύνοντας ένα πρόβλημα αντίστροφης ενισχυτικής μάθησης στοχεύουμε στην εύρεση της συνάρτησης ανταμοιβής δοθέντος της αντίστοιχης βελτιστης πολιτικής $\pi^*$. Με άλλα λόγια, δοθέντος της πολιτικής $\pi^*$ και του μοντέλου $\{S, A, P, \gamma\}$, θέλουμε να υπολογίσουμε την συνάρτηση ανταμοιβής $r$ τέτοια ώστε η πολιτική $\pi^*$ να είναι βελτιστη για τη Μαρκοβιανή Διαδικασία Απόφασης $\{S,\ A,\ P,\ r,\ \gamma\}$. Από τις σχέσεις 2.6, 2.11 έχουμε :

$$R^{\alpha}_{ss'} = Q^*(s,\alpha) - \gamma \sum_{s'} P^{\alpha}_{ss'} \sum_{\alpha'} \pi^*(s',\alpha') Q(s',\alpha') \tag{2.29}$$

Αν η εξίσωση Bellman ορίζει τη βέλτιστη συνάρτηση αξίας κατάστασης-ενέργειας δοθέισας της συνάρτηση ανταμοιβής η (2.29) ορίζει το αντίστροφο, δηλαδή τη συνάρτηση ανταμοιβής δοθέισας της $Q$ και ονομάζεται αντίστροφη *Bellman* εξίσωση (*inverse Bellman equation*). Τόσο η εξίσωση *Bellman* αλλά και η αντίστροφη εξίσωση *Bellman* ορίζουν μία προς μία αντιστοιχία μεταξύ της συνάρτησης ανταμοιβής και της συνάρτησης κατάστασης-ενέργειας $Q$. Έχοντας δηλαδή οποιαδήποτε $Q$ υπάρχει μια αντιστοιχία σε μια συνάρτηση ανταμοιβής $r$ τέτοια ώστε το $Q$ να είναι η βέλτιση συνάρτηση αξίας κατάστασης-ενέργειας που σχετίζεται με την $r$.

Οταν ο χώρος καταστάσεων είναι διακριτός, έστω $S$, κι $A = (\alpha_1 \cdots \alpha)$, μια πολιτική $\pi$ που δινεται από $\pi(s) \equiv \alpha_1$ είναι βέλτιστη αν και μόνο αν είναι καλύτερη από κάθε άλλη πολιτική $(\alpha_2, \cdots, \alpha_k)$. Η εξίσωση *Bellman* για την συνάρτηση αξίας κατάστασης γράφεται σε αλγεβρική μορφή ως :

$$V^{\pi} = R + \gamma P_{\alpha_1} V^{\pi} \tag{2.30}$$

$$V^{\pi} = (I - \gamma P_{\alpha_1})^{-1} R \tag{2.31}$$

Από την 2.12 έχουμε ότι $\alpha_1 \equiv (s) \in \arg\max \sum_{s'} P_{s\alpha}(s')V(s') \ \ \forall s \in S$. Οπότε

$$\begin{aligned}
&\alpha_1 \geq \alpha \\
&\Leftrightarrow \sum_{s'} P_{s\alpha_1}(s')V^{\pi}(s') \geq \sum_{s'} P_{s\alpha}(s')V^{\pi}(s') \ \ \forall s \in S, \alpha \in A \\
&\Leftrightarrow P_{\alpha_1}V^{\pi} \geq P_{\alpha}V^{\pi} \ \ \forall \alpha \in A \setminus \alpha_1 \\
&\Leftrightarrow P_{\alpha_1}(-\gamma P_{\alpha_1})^{-1}R \geq P_{\alpha}(I - \gamma P_{\alpha_1})^{-1}R \ \ \forall \alpha \in A \setminus \alpha_1
\end{aligned} \tag{2.32}$$

Προκύπτει επομένως ένα πρόβλημα βελτιστοποίηση με περιορισμούς για την εύρεση της συνάρτησης ανταμοιβής $R$:

$$(P_{\alpha_1} - P_{\alpha})(I - \gamma P_{\alpha_1})^{-1}R \succ 0 \tag{2.33}$$

Μια προφανής λύση είναι $R = 0$. Υπάρχουν όμως και πολλές επιλογές αν το $R$ δεν είναι μηδέν που ικανοποιούν την σχέση 2.34. Από όλες αυτές τις επιλογές προτιμουμε αυτή που μεγιστοποιεί το άθροισμα της διαφοράς της συνάρτησης αξίας κατάστασης-ενέργειας της βέλτιστης πολιτικής με αυτή της επόμενης καλύτερης ενέργειας. Το πρόβλημα βελτιστοποίησης μετατρεπεται επομένως σε:

$$\sum_{s \in S} \left( Q(s, \alpha_1) - \max_{\alpha \in A \setminus \alpha_1} Q(s, \alpha) \right) \tag{2.34}$$

Για την ομαλοποίηση της λύσης και για υπολογιστικούς λόγους χρησιμοποιείται ένα τελεστής ομαλοποίησης $\lambda$ ώστε οι λύσεις για την συνάρτηση ανταμοιβής να απομακρυνθούν από την τιμή 0. Με τη χρήση αυτού του τελεστή το πρόβλημα βελτιστοποίησης γίνεται:

$$\sum_{i=1}^{N} \min_{\alpha \in (\alpha_2 \cdots \alpha_k)} ((P_{\alpha_1}(i) - P_{\alpha}(i))(I - \gamma P_{\alpha_1})^{-1}R) - \lambda ||R|| \tag{2.35}$$

με τους περιορισμούς $(P_{\alpha_1} - P_{\alpha})(I - \gamma P_{\alpha_1})^{-1}R \succ 0 \ \forall \alpha \in A \setminus \alpha_1$ και $|R_i| \leq R_{max}, i = 1, \cdots, N$, όπου $P_{\alpha}(i)$ είναι η i γραμμή του πίνακα $P_{\alpha}$

Όταν το χώρος καταστάσεων είναι άπειρος τότε η συνάρτηση ανταμοιβής μπορεί να γραφεί ως γραμμικός συνδιασμός των παραμέτρων $u$ και των $k$ χαρακτηριστικών δηλαδή:

$$R(s, a) = u_1\phi_1(s, \alpha) + u_2\phi_2(s, \alpha) + \cdots + u_k\phi_k(s, \alpha) \tag{2.36}$$

Αν θεωρήσουμε τη συνάρτηση αξίας κατάστασης της πολιτικής $\pi$ ως $V_i$ και η συνάρτηση ανταμοιβής είναι $R = \phi_i$ εξαιτίας της γραμμικότητας έχουμε:

$$V^{\pi} = u_1 {V_1}^{\pi} + \cdots + u_k {V_k}^{\pi} \tag{2.37}$$

κι από την 2.12 προκύπτει το πρόβλημα βελτιστοποίησης

$$E_{s' \sim P_{s\alpha_1}} \left[ V^{\pi}(s') \right] \geq E_{s' \sim P_{s\alpha}} \left[ V^{\pi}(s') \right] \tag{2.38}$$

Για να μην επιτραπούν μεγάλες τιμές στις παραμέτρους $u$ και ώστε να γίνει το πρόβλημα πιο γενικό χρησιμοποιεί περιορισμους.

$$\sum_{i=1}^{N} \min_{\alpha \in (\alpha_2 \cdots \alpha_k)} \left\{ p(E_{s' \sim P_{s\alpha_1}} \left[ V^{\pi}(s') \right] - E_{s' \sim P_{s\alpha}} \left[ V^{\pi}(s') \right]) \right\} \tag{2.39}$$

με $|u| \leq 1, i = 1, \cdots, k$. Όπου $p$ μια συνάρτηση η οποία ισούται με $p(x) = x$ όταν το $x \geq 0$ αλλιώς $p(x) = 2x$.

# Κεφαλαιο 3

# Η Πλατφορμα Delta Berenike



## 3.1 Εισαγωγή

Οι επιπλέουσες πλατφόρμες χρησιμοποιούνται ευρέως στην υπεράκτια βιομηχανία πετρελαίου, ως βοηθητικές πλατφόρμες επισκευών σε ναυπηγεία καθώς και σε πλήθος άλλων εφαρμογών. Η πλατφόρμα *Delta Berenike* κατασκευάστηκε για να χρησιμοποιηθεί ως πλωτή βάση εξυπηρέτησης στην κατασκευή του τηλεσκοπίου "NESTOR". Αυτή η πλατφόρμα, κατά τη διάρκεια της λειτουργίας της στην επιφάνεια του νερού, θα έπρεπε να διατηρεί τη θέση και τον προσανατολισμό της μέσα σε μια προκαθορισμένη περιοχή γύρω από το σημείο κατασκευής του τηλεσκοπίου. Είναι εφοδιασμένη με κατάλληλο σύστημα κίνησης που της παρέχει τη δυνατότητα οδήγησης και αυτόνομου δυναμικού ελέγχου της θέσης της, ώστε να μπορεί να αντιμετωπίσει ορισμένης έντασης κύματα και θαλάσσια ρεύματα κατά τη διάρκεια εργασιών εξυπηρέτησης του "NESTOR". Σημαντικά ακόμα προβλήματα στο σύστημα ελέγχου προέρχονται και από τη μη πλήρη γνώση του σύνθετου υδροδυναμικού μοντέλου, του μοντέλου των επενεργητών καθώς και την ανακρίβεια που

προέρχεται από την ανατροφοδότηση των μεταβλητών κατάστασης (τρέχουσα θέση, προσανατολισμός και ταχύτητα) από τους αισθητήρες κίνησης. Στη συνέχεια του κεφαλαίου αυτού θα παρουσιαστούν η γεωμετρία της πλατφόρμας, η κινηματική και η δυναμική της καθώς και οι διάφορες δυνάμεις που επενεργούν πάνω της.

## 3.2 Φυσικά Χαρακτηριστικά της Πλατφόρμας

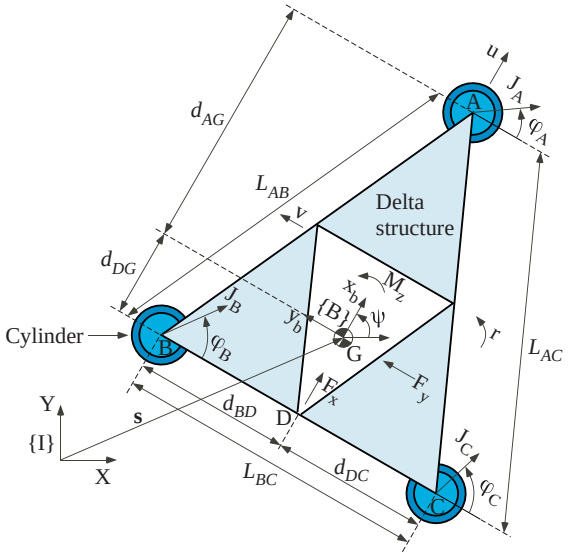
Η πλατφόρμα *Delta Berenike*, σχήμα (3.1) αποτελείται από μια τριγωνική κατασκευή η οποία στηρίζεται πάνω σε τρεις διπλούς κυλίνδρους (ομοαξονικούς και διαφορετικής διαμέτρου), ένας σε κάθε κορυφή της όπως απεικονίζεται στο σχήμα (3.2).


Σχήμα 3.1: Η τριγωνική θαλάσσια πλατφόρμα Delta Berenike.

Το επίπεδο του τριγώνου είναι παράλληλο με το επίπεδο της θάλασσας. Οι κύλινδροι οι οποίοι είναι μισοβυθισμένοι μέσα στο νερό παρέχουν την απαραίτητη άνωση έτσι ώστε, η κατασκευή να βρίσκεται εκτός του νερού. Τη δυνατότητα κίνησης της πλατφόρμας τη δίνουν οι τρεις αντλίες-τζετ νερού που βρίσκονται στο κάτω μέρος του κάθε κυλίνδρου και είναι πλήρως βυθισμένες μέσα στο νερό. Μια μηχανή *diesel* κινεί κάθε αντλία, ενώ ένας ηλεκτρο-υδραυλικός κινητήρας περιστρέφει το τζετ παρέχοντας κατευθυνόμενη πρόωση.

### 3.2.1 Η γεωμετρία της πλατφόρμας

Η κυρίως κατασκευή της πλατφόρμας έχει το σχήμα ενός ισοσκελούς τριγώνου με τις ίσες πλευρές $L_{AB} = L_{AC}$ και με μήκος βάσης $L_{BC}$, σχήμα 3.2.




Σχήμα 3.2: Δισδιάστατη αναπαράσταση της πλατφόρμας.

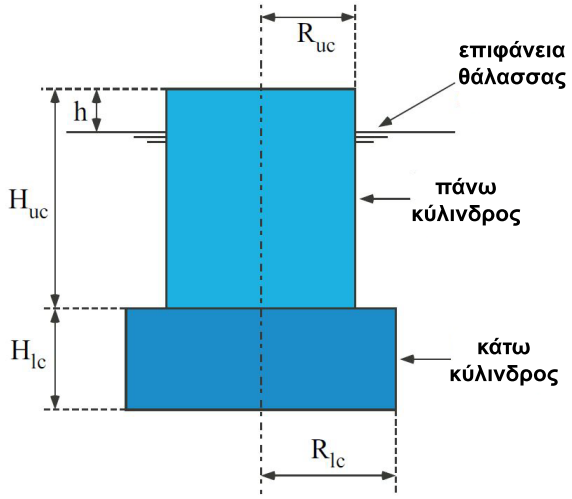
Το κέντρο μάζας (ΚΜ) της κατασκευής συμπίπτει με το σημείο G κατά μήκος του άξονα συμμετρίας σε απόσταση $d_{AG}$ από την κορυφή $A$. Όλα τα γεωμετρικά χαρακτηριστικά της πλατφόρμας περιγράφονται στον ακόλουθο πίνακα.

| **Σύμβολισμός** | **Μέγεθος** | **Μονάδες** | **Περιγραφή** |
|---|---|---|---|
| m | $425 \times 10^3$ | $kg$ | συνολική μάζα πλατφόρμας |
| $I_{zz}$ | $25.73 \times 10^7$ | $Nms^2$ | μαζική ροπή αδράνειας |
| $L_{AB}$ | 44 | $m$ | πλευρά τριγώνου |
| $L_{AC}$ | 44 | $m$ | πλευρά τριγώνου |
| $L_{BC}$ | 36.483 | $m$ | βάση τριγώνου |
| $L_{AD}$ | 40.041 | $m$ | διάμεσος τριγώνου |
| $d_{AG}$ | 34.796 | $m$ | απόσταση AG |
| $d_{DG}$ | 5.245 | $m$ | απόσταση DG |
| $d_{BD}$ | 18.2415 | $m$ | απόσταση BD |
| $d_{DC}$ | 18.2415 | $m$ | απόσταση DC |

Τα γεωμετρικά χαρακτηριστικά των τριών κυλίνδρων που είναι όμοιοι, φαίνονται στο Σχημα 3.3. Η ποσότητα $h$ είναι μεταβλητή και συμβολίζει το ύψος του άνω κυλίνδρου που βρίσκεται πάνω από την επιφάνεια του νερού. Στο ίδιο σχήμα,

φαίνεται ότι ο κάτω κύλινδρος είναι πλήρως βυθισμένος. Στο κάτω μέρος αυτού του κυλίνδρου είναι προσαρμοσμένα τα τζετ νερού που είναι υπεύθυνα για την κίνηση της πλατφόρμας. Στον πίνακα που ακολουθεί παρουσιάζονται τα χαρακτηριστικά των κυλίνδρων:

| **Σύμβολισμός** | **Μέγεθος** | **Μονάδες** | **Περιγραφή** |
|---|---|---|---|
| $R_{uc}$ | 2.2 | $m$ | ακτίνα άνω κυλίνδρου |
| $R_{lc}$ | 3.4 | $m$ | ακτίνα κάτω κυλίνδρου |
| $H_{uc}$ | 6.3 | $m$ | ύψος άνω κυλίνδρου |
| $H_{cl}$ | 3.0 | $m$ | ύψος κάτω κυλίνδρου |




Σχήμα 3.3: Πλάγια όψη της κατασκευής των διπλών κυλίνδρων.

## 3.3 Η κινηματική

Η κινηματική διαπραγματεύεται τη μελέτη της κίνησης των σωμάτων χωρίς όμως να ασχολείται με τις δυνάμεις ή τις ροπές που την προκαλούν. Στο κεφάλαιο αυτό μελετάμε την κινηματική όπως εφαρμόζεται στη πλατφόρμα *Delta Berenike*. Για να περιγραφεί η κινηματική, χρησιμοποιούνται δύο συστήματα συντεταγμένων, το αδρανειακό σύστημα συντεταγμένων $\Sigma\Sigma\{I\}$ και το σωματόδετο $\Sigma\Sigma\{B\}$, σχήμα (3.2). Όπως παρατηρείτε, η αρχή του $\Sigma\Sigma\{B\}$ συμπίπτει με το ΚΜ. Ο άξονας $x_b$ συμπίπτει με τον άξονα συμμετρίας της πλατφόρμας, ο $y_b$ δείχνει αριστερά και ο $z_b$ προς τα πάνω (εκτός του επίπεδου του σχήματος). Τότε, οι εξισώσεις κινηματικής για την επίπεδη κίνηση είναι:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ r \end{bmatrix} \Rightarrow {}^{I}\mathbf{x} = {}^{I}\mathbf{R}_{B}{}^{B}\mathbf{v} \tag{3.1}$$

Στην εξίσωση (3.1), τα $x$ και $y$ παριστάνουν τις αδρανειακές συνιστώσες του ΚΜ και το $\psi$ είναι ο προσανατολισμός του $\Sigma\Sigma\{B\}$ ως προς το $\Sigma\Sigma\{I\}$. Τα $u$, $v$ είναι η πρόσθια και η πλάγια ταχύτητα αντίστοιχα, εκφρασμένες στο σωματόδετο σύστημα συντεταγμένων και r είναι η γωνιακή ταχύτητα της πλατφόρμας.

## 3.4 Η δυναμική

Η δυναμική ασχολείται με την εξαγωγή και τη μελέτη του δυναμικού μοντέλου μιας ρομποτικής κατασκευής. Το δυναμικό μοντέλο συνίσταται από τις διαφορικές εξισώσεις που περιγράφουν αναλυτικά τη σχέση ανάμεσα στις δυνάμεις/ροπές των επενεργητών και την γραμμική και γωνιακή επιτάχυνση της κατασκευής μέσα στο περιβάλλον. Κατά τη διάρκεια πλοήγησης της θαλάσσιας πλατφόρμας πάνω της επενεργούν τρεις διαφορετικοί τύποι δυνάμεων/ροπών, οι οποίοι εκφράζονται ως προς το κέντρο μάζας της πλατφόρμας. Ένα πρώτο είδος δυνάμεων που δρουν στην πλατφόρμα είναι οι δυνάμεις που προκαλούνται από τους κινητήρες τις πλατφόρμας. Επιπλέον, η πλοήγηση της πλατφόρμας επηρεάζεται από τις υδροδυνάμεις ενώ παράλληλα δέχεται επιδράσεις και από περιβαλλον.

### 3.4.1 Δυνάμεις από τους κινητήρες

Οι κινητήρες της πλατφόρμας όπως έχει ήδη αναφερθεί, έχουν τη δυνατότητα να παρέχουν κατευθυνόμενη πρόωση και συνεπώς μεγαλύτερη ευελιξία στο σχεδιασμό ελέγχου. Τα μέτρα των τριών προώσεων συμβολίζονται με $J_A$, $J_B$ και $J_C$ , ενώ $\phi_A$, $\phi_B$ και $\phi_C$ συμβολίζουν τις αντίστοιχες μεταβλητές περιστροφής των κινητήρων. Οι προώσεις αυτής της διαμόρφωσης παρέχουν δυναμεις ελέγχου στους $x_b$ και $y_b$ άξονες, $F_x$ και $F_y$ ως προς το ΚΜ και τη ροπή ελέγχου $M_z$ ως προς τον $z_b$ άξονα, σύμφωνα με τον ακόλουθο μετασχηματισμό:

$${}^{B}\mathbf{n}_c = [F_x, F_y, M_z]^T = \mathbf{B}\,{}^{B}\mathbf{f}_c \tag{3.2}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & -d_{AG} \\ 1 & 0 & -d_{DC} \\ 0 & -1 & d_{DG} \\ 1 & 0 & d_{DC} \\ 0 & -1 & d_{DG} \end{bmatrix}^T \quad , \quad {}^B\mathbf{f}_c = \begin{bmatrix} J_A \sin\phi_A \\ J_A \cos\phi_A \\ J_B \sin\phi_B \\ J_B \cos\phi_B \\ J_C \sin\phi_C \\ J_C \cos\phi_C \end{bmatrix} \tag{3.3}$$

όπου ${}^B\mathbf{n}_c$ είναι το διάνυσμα ελέγχου δυνάμης και ροπής. Η επιθυμητή πρόωση και κατεύθυνση του κάθε κινητήρα υπολογίζεται σύμφωνα με την παρακάτω εξίσωση:

$$J_i = \sqrt{(f_i \sin\phi_i)^2 + (f_i \cos\phi_i)^2} \tag{3.4}$$

$$\phi_i = \arctan(f_i \sin\phi_i, f_i \cos\phi_i) \tag{3.5}$$

όπου $i = A, B, C$.

Για την εύρεση του διανύσματος ${}^B\mathbf{n}_c$ απαιτείται η αντιστροφή του πίνακα $\mathbf{B}$. Ο πίνακας $\mathbf{B}$ όμως είναι τετραγωνικός επομένως χρησιμοποιούμε την μέθοδος της ψευδοαντιστροφής. Ο συγκεκριμένος πίνακας δεν είναι τετραγωνικός επειδή το σύστημά μας είναι *overactuated*, δηλαδή έχουμε περισσότερες μεταβλητές ελέγχου (6, 3 $\phi_i$ + 3 $J_i$) από ελεγχόμενους βαθμούς ελευθερίας (3, $x, y, \psi$).

### 3.4.2 Υδροδυνάμεις

Οι υδροδυνάμεις είναι το αποτέλεσμα της κίνησης των κυλίνδρων στο νερό. Είναι δηλαδή η δύναμη που εμφανίζεται ως πρόσθετη μάζα εξαιτίας της επιτάχυνσης του νερού που περιβάλλει τον επιταχυνόμενο κύλινδρο και είναι γραμμική συνάρτηση της επιτάχυνσης του κυλίνδρου. Αντίθετα η δύναμη της αντίστασης του νερού είναι τετραγωνική συνάρτηση της σχετικής ταχύτητας μεταξύ του κάθε κυλίνδρου και του νερού. Αυτές οι δυνάμεις μοντελοποιούνται σύμφωνα με την εξίσωση του *Morison* και σαν παράδειγμα καταγράφουμε την ασκούμενη δύναμη στον κύλινδρο του σημείου του τριγώνου, εκφρασμένη στο σωματόδετο $\Sigma\Sigma\{B\}$ :

$$\begin{aligned} {}^B\mathbf{f}_{h,A} = C_a \pi \rho_w [R_{uc}^2 (H_{uc} - h) + R_{lc}^2 H_{lc}](-{}^B\mathbf{a}_A) + \\ C_d \rho_w [R_{uc}(H_{uc} - h) + R_{lc} H_{lc}] \\ \|({}^B\mathbf{v}_{wat} - {}^B\mathbf{v}_A)\|({}^B\mathbf{v}_{wat} - {}^B\mathbf{v}_A) \end{aligned} \tag{3.6}$$

με το $\rho_w$ να εκφράζει την πυκνότητα του νερού, $C_a$ είναι ο συντελεστης πρόσθετης μάζας και $C_d$ είναι ο συντελεστής του νερού. Το ${}^B\mathbf{v}_A$ συμβολίζει την ταχύτητα του

κυλίνδρου , το ${}^{B}\mathbf{a}_{A}$ την επιτάχυνση του εκφρασμένες ως προς το σωματόδετο $\{\}$ και το ${}^{B}\mathbf{v}_{wat}$ την ταχύτητα του νερού. Οι παράμετροι $h$, $R_{uc}$, $H_{uc}$, $R_{lc}$, και $H_{lc}$ συμβολίζουν το ύψος κάθε κυλίνδρου πάνω από την επιφάνεια της θάλασσας και την ακτίνα αλλα και το ύψος του πάνω και κάτω τμήματος του κυλίνδρου αντίστοιχα. Έτσι οι υδροδυνάμεις που ασκούνται στο σημείο $A$ και δίνονται από το (3.6) έχουν ως αποτέλεσμα μια δύναμη και μια ροπή που ασκούνται στο ΚΜ της πλατφόρμας σύμφωνα με την εξίσωση:

$$ {}^{B}\mathbf{q}_{h,A} = [{}^{B}\mathbf{f}_{h,A}^{T}, ({}^{B}\mathbf{s}_{A/G} \times^{B} \mathbf{f}_{h,A})^{T}]^{T} \tag{3.7} $$

όπου ${}^{B}\mathbf{s}_{A/G}$ είναι η θέση του σημείου  ως προς το G εκφρασμένη στο σωματόδετο $\Sigma\Sigma\{B\}$. Όλες αυτές οι δυνάμεις από κάθε κορυφή της πλατφόρμας όπου βρίσκεται ο κάθε κινητήτας είναι τετραγωνικες συνάρτησεις της ταχύτητας και μπορουμε να τις συνοψίσουμε με το ακόλουθο διάνυσμα:

$$ {}^{B}\mathbf{q} = [f_x, f_y, n_z]^{T} \tag{3.8} $$

### 3.4.3 Περιβαλλοντικές διαταραχές

Εκτός από τις δυνάμεις που αναφέρθηκαν, η κίνηση της πλατφόρμας επηρεάζεται και από άλλες δυνάμεις που ασκούνται πανω της από το περιβάλλον. Αυτές οι περιβαλλοντικές διαταραχές αφορούν την επιρροή του ανέμου πάνω στην πλατφόρμα, τα θαλάσσια ρεύματα καθώς και τους κυματισμούς της θάλασσας.

#### 3.4.3.1 Δυνάμεις από τον άνεμο

Οι δυνάμεις που προκαλούνται από τον άνεμο υπολογίζονται σύμφωνα με τις παρακάτω εξισώσεις:

$$ \begin{aligned} f_{x,wind} &= 0.5\ C_X(\gamma_R)\rho V_R^2 A_T \\ f_{y,wind} &= 0.5\ C_Y(\gamma_R)\rho V_R^2 A_L \\ n_{z,wind} &= 0.5\ C_T(\gamma_R)\rho V_R^2 A_L L \end{aligned} \tag{3.9} $$

$$ {}^{B}\mathbf{q}_{wind} = [f_{x,wind}, f_{y,wind}, n_{z,wind}]^{T} \tag{3.10} $$

όπου $C_X$ και $C_Y$ είναι συντελεστές των δυνάμεων. Αυτοί οι συντελεστές είναι συναρτήσεις της σχετικής γωνίας $\gamma_R$ μεταξύ του ανέμου και του προσανατολισμού της

πλατφόρμας. Το $\rho$ εκφράζει την πυκνότητα του αέρα ενώ τα  και $_L$ είναι οι εγκάρσιες και πλευρικές προβολές και το $L$ το ολικό μήκος της πλατφόρμας. Η σχετική ταχύτητα του ανέμου συμβολίζεται με $V_R$ και μετριέται σε κόμβους. Το μέτρο της ταχύτητας του ανέμου συμβολίζεται με $\upsilon_t$ και η μέγιστη τιμή της είναι $\upsilon_t \leq 7.9m/s$, δηλαδή 15 κόμβοι ή 4 μποφόρ.

#### 3.4.3.2 Δυνάμεις από τα κύματα και θαλλάσια ρεύματα

Η δυνάμεις των κυμάτων εξαρτώνται από την ένταση του αέρα κάθε χρονική στιγμή και εισάγωνται στο μοντέλο της πλατφόρμας εκφράζοντας τη δυναμική εξίσωση, ως προς την σχετική ταχύτητα ανάμεσα στην πλατφόρμα και την ταχύτητα του νερού. Παρόμοια αντιμετωπίζονται και τα θαλάσσια ρεύματα.

Χρησιμοποιώντας τους παραπάνω υπολογισμούς εξάγουμε τις εξισώσεις επίπεδης κίνησης της πλατφόρμας εκφρασμένες στο σωματόδετο $\Sigma\Sigma\{B\}$ :

$$\mathbf{M}^B\dot{\mathbf{v}} = {}^B\mathbf{q} + {}^B\mathbf{q}_{wind} + {}^B\mathbf{n}_c \tag{3.11}$$

$$\mathbf{M} = \begin{bmatrix} m - 3m_a & 0 & 0 \\ 0 & m - 3m_a & 0 \\ 0 & 0 & m_{33} \end{bmatrix} \tag{3.12}$$

$$m_{33} = I_{zz} - (d_{AG}^2 + 2d_{BD}^2 + 2d_{DG}^2)m_a \tag{3.13}$$

όπου $m$ είναι η μάζα της πλατφόρμας, $m_a$ είναι η επιπρόσθετη μάζα και $I_{zz}$ η μαζική ροπή αδράνειας ως προς τον άξονα $z_b$. Το $q_{wind}$ περιέχει τις δυνάμεις απο τα κύματα και τα θαλάσσια ρεύματα.

## 3.5 Θόρυβος Μετρήσεων

Με σκοπό την ολοκληρωμένη μοντελοποίηση του συστήματος οι μετρήσεις της θέσης της πλατφόρμας, κατά τη διάρκεια των προσομοιώσεων, περιέχουν και θόρυβο μετρήσεων σύμφωνα με τις συσκευές *GPS*. Αυτές οι μετρήσεις είναι πραγματικές και λήφθηκαν αφαιρώντας τη μέση τιμή των μετρήσεων έπειτα από εικοσιτετράωρη λειτουργία των συσκευών *GPS*.

# Κεφαλαιο 4

# Κατασκευη ενος ευφυη πρακτορα ενισχυτικης μαθησης για την πλοηγηση της θαλασσιας ρομποτικης πλατφορμας



## 4.1 Ορισμός του προβλήματος

Στην παρούσα διατριβή το ασχοληθήκαμε με το πρόβλημα της κατασκευης ενός ευφυή πράκτορα ενισχυτικής μάθησης για την πλοήγηση της θαλάσσιας ρομποτικής πλατφόρμας *Delta Berenike*. Στόχος της διατριβής αποτέλεσε η μοντελοποίηση της κίνησης και της αποφυγής εμποδίων της ρομποτικής πλατφόρμας ώστε να επιτυχγάνεται η αυτόνομη πλοήγηση της μέσα σε ένα άγνωστο περιβάλλον. Ξεκίνωντας από μια αρχική θέση ο πράκτορας προσπαθεί να επιλέξει την βέλτιστη διαδρομή αποφεύγοντας πιθανά εμπόδια με σκοπό την προσέγγιση μιας συγκεκριμένης θέσης (*docking location*). Κατά τη διάρκεια της αλληλεπίδρασης με το άγνωστο περιβάλλον επιδρούν περιβαλλοντικές διαταραχές (άνεμος, θαλάσσια ρεύματα, κυματα) που

δυσκόλευου τη κίνηση της πλατφόρμας και διαμορφώνουν ένα στοχαστικό περιβάλλον.

## 4.2 Ορισμός χώρου καταστάσεων και ενεργειών

### 4.2.1 Ορισμός χώρου καταστάσεων

Η αναπαράσταση του χώρου καταστάσεων αποτελεί ένα αναποσπαστο κομμάτι των μεθόδων που βασίζονται στην ενισχυτική μάθηση. Η εύρεση μις αποτελεσματικής αναπαράστασης του χώρου καταστάσεων είναι μια δύσκολη και συνάμα απαιτητική διαδικασία. Η αναπαράσταση θα πρέπει να περιελαμβάνει όλη εκείνη την απαραίτητη πληροφορία που χρειάζεται ένας πράκτορας ώστε να φθάσει στην ανακάλυψη της βέλτιστης πολιτικής. Θα πρέπει να σημειωθεί ότι η πολυπλοκότητα των αλγορίθμων ενισχυτικής μάθησης συνδέεται άμεσα με το μέγεθος του χώρου καταστάσεων. Για παράδειγμα, ένας τεράστιος χώρος καταστάσεων θα αυξήσει τόσο τη πολυπλοκότητα του αλγορίθμου όσο και τις απαιτήσεις σε φυσική μνήμη καθώς και σε υπολογιστική ισχύ. Την ίδια στιγμή, ένα μεγάλο εύρος πληροφορίας μπορεί να οδηγήσει σε μείωση της γενικευτικής ικανότητας του εκάστοτε αλγορίθμου καθιστώντας τον μη πρακτικό σε άγνωστα σε αυτόν περιβάλλοντα.

Η πλοήγηση μιας θαλάσσιας ρομποτικής πλατφόρμας αποτελεί ένα δύσκολο τομέα για να δοκιμάσει κανείς έναν ευφυή πράκτορα. Η αναπαράσταση του χώρου καταστάσεων είναι το σημαντικότερο γεγονός, μιας και αποτελέι σημαντικό ρόλο σε ένα σύστημα μοντελοποίησης, αναγνώρισης και προσαρμοστικού ελέγχου. Σε κάθε βήμα ο πράκτορας θα πρέπει να πάρει μια σωστή απόφαση σύμφωνα με ότι παρατηρεί γύρω του. Η περιγραφή του χώρου καταστάσεων στο πρόβλημα που καλούμαστε να αντιμετωπίσουμε αποτελείται από:

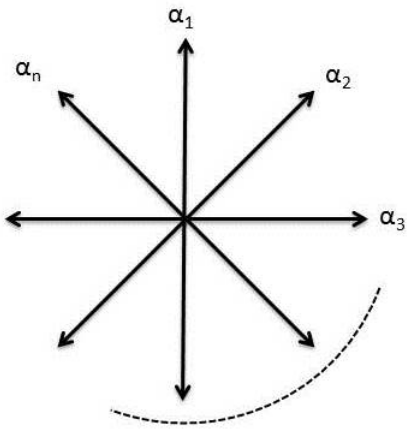
- την σχετική θέση (συντεταγμένες) της πλατφόρμας στο εκάστοτε χάρτη,
- την γωνία που περιγράφει τον προσανατολισμό της,
- τις γραμμικές ταχύτητες καθώς και την γωνιακή της.

Για την απλοποίηση κατά μία έννοια αυτού επιλέξαμε να χρησιμοποιήσουμε μονο τη σχετική θέση της πλατφόρμας αλλά και την γωνία που περιγράφει τον προσανατολισμό της.

Έχοντας τη σχετική θέση της πλατφόρμας κάθε χρονική στιγμή το επόμενο χαρακτηριστικό του χώρου καταστάσεων που κληθήκαμε να ορίσουμε ήταν η γωνία που περιγράφει τον προσανατολισμό της πλατφόρμας. Ως γωνία προσανατολισμού της πλατφόρμας θέσαμε τη γωνία την οποία έχει ο άνεμος σε κάθε χρονική στιγμή. Κάνοντας αυτή την απλοποίηση επιτυγχάνουμε την μείωση του χώρου καταστάσεων μειώνοντας ταυτόχρονα και την αντίσταση του ανέμου πάνω στην κατασκευή της πλατφόρμας.

### 4.2.2 Ορισμός χώρου ενεργειών

Ένα εξίσου σημαντικό κομμάτι των μεθόδων ενισχυτικής μάθησης είναι αυτό του ορισμού του χώρου ενεργειών $A = \{\alpha_i | i = 1, \cdots, M\}$. Ως χώρο ενεργειών στο πρόβλημά μας ορίσαμε το μέτρο αλλά και τον προσανατολισμό που μπορει να χρησιμοποιήσουν τα τζετ-νερού της πλατφόρμας για να μπορέσουν με την περιστροφή τους να κατευθύνουν την πλατφόρμα. Το τιμή του μέτρου της δύναμης τη θεωρήσαμε σταθερή. Για την επιλογή των γωνιών κατα την πειραματική αξιολόγηση της μεθόδου μελετήσαμε διάφορες γωνίες για το χώρο των ενεργειών. Παρόλα αυτά το



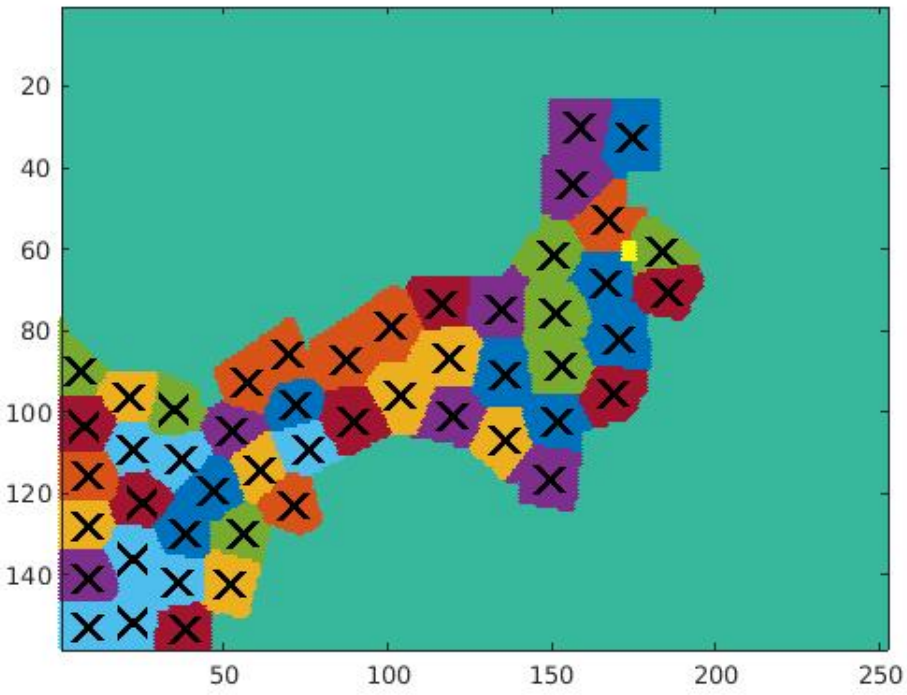

Σχήμα 4.1: Αναπαράσταση των διαφόρων γωνιών

μέτρο της δύναμης θα μπορούσε να αποτελείται από περισσότερες από μία τιμές. Με αυτό τον τρόπο η εύρεση του βέλτιστου μονοπατιού θα σχετιζόταν με την επιλογή της κατάλληλης τιμής του μέτρου. Αυτό θα είχε ως αποτέλεσμα έμμεσα την εξοικονόμιση ενέργειας στην πλοήγηση της ρομποτικής πλατφόρμας.

## 4.3 Επιλογή των χαρακτηριστικών $\phi$

Η επιλογή των συναρτήσεων βάσης αποτελεί ένα πολύ σημαντικό στάδιο της προτεινόμενη μέθοδο. Σε κάθε κατάσταση $s$ αντιστοιχεί ένα διάνυσμα χαρακτηριστικών $\phi(s) = \{\phi_i(s)|i = 1, \cdots, k\}$ τα οποία πρέπει να επιλεγούν με τέτοιο τρόπο ώστε να κωδικοποιούν ιδιότητες των καταστάσεων αλλά και των ενεργειών σχετικά με την κατάλληλη επιλογή των συναρτήσεων αξιών κατάστασης-ενέργειας $Q$.

Στην προτεινόμενη μεθοδολογία που θα περιγραφεί στην συνέχεια χρησιμοποιήσαμε $k$ συναρτήσεις ακτινικής βάσης. Για να υπολογίσουμε τα χαρακτηριστικά της κάθε συνάρτησης $\phi_i(s)$, δηλαδή το κέντρο της $c_i$ αλλά και το πλάτος της $\sigma_i$ ακολουθήσαμε τον τρόπο της κωδικοποίησης πλακιδίων. Διαμερίσαμε το χώρο στον οποίο μπορεί να κινηθεί η πλατφόρμα ομοιόμορφα σε $k$ μη επικαλυπτόμενες περιοχές και δημιουργήσαμε το διάνυσμα των $k$ κέντρων. Αφού μελετήσαμε την συμπεριφορά της μεθόδου για διάφορες τιμές του $k$, καταλήξαμε στην τιμή $k = 50$. Μια διαμέριση του χώρου σε 50 κέντρα $c_i$ απεικονίζεται στο Σχήμα 4.2. Όσον αφορά το πλατος




Σχήμα 4.2: Διαμερισμός του χώρου με 50 κέντρα

$\sigma_i$ της κάθε συνάρτησης επιλέξαμε να είναι κοινό για όλες τις συναρτήσεις βάσης. Έπειτα από μελέτη της συμπεριφοράς της συγκεκριμένης παραμέτρου ορίσαμε την

τιμή $\sigma_i = 0.1$.

## 4.4 Η προτεινόμενη μέθοδος εύρεσης βελτιστης πολιτικής και συνάρτησης ανταμοιβής

Η προτεινόμενη μεθοδολογίας περιελαμβάνει την ταυτόχρονη εύρεση της πολιτικής αλλά και της συνάρτησης ανταμοιβής. Όπως αναφέραμε και σε προηγούμενη ενότητα όταν ο χώρος καταστάσεων είναι πολύ μεγάλος ή άπειρος τότε χρησιμοποιείται μια προσέγγιση της συνάρτησης κατάστασης ή κατάστασης ενέργειας.΄Ενα μαθηματικό πλαίσιο που χρησιμοποιείται είναι το γραμμικό μοντέλο στο οποίο η συνάρτηση αξίας κατάστασης-ενέργειας $Q$ εκτιμάται χρησιμοποιώντας $k$ γραμμικούς συντελεστές $w$ και $k$ χαρακτηριστικά $\phi$, δηλαδή $Q(s,\alpha) = \phi(s,\alpha)^\top w$.

Αντίστοιχα η συνάρτηση ανταμοιβής μπορεί να προσεγγιστεί με ανάλογο τρόπο με την χρήση $k$ αγνώστων παραμέτρων $u$ και $k$ των χαρακτηριστικών $\phi$. Οπότε η συνάρτηση ανταμοιβής μπορεί να γραφεί ως:

$$R(s,\alpha) = \phi(s,\alpha)^\top u \tag{4.1}$$

### 4.4.1 Εκτίμηση παραμέτρων της μεθόδου

Για να εκτιμήσουμε τη συνάρτηση αξίας κατάστασης-ενέργειας χρησιμοποιούμε ένα σύνολο δειγμάτων $D = \{s_i, a_i, r_i, s_i' | i = 1, \ldots, n\}$ τα οποία μπορεί να ληφθούν είτε με τυχαίο τρόπο είτε απο μια πραγματική επεισοδιακή διαδικασία. Οι πίνακες $\Phi, \Phi'$ είναι της μορφής:

$$\Phi = \begin{pmatrix} \phi(s_1,\alpha_1)^\top \\ \vdots \\ \phi(s,\alpha)^\top \\ \vdots \\ \phi(s_n,\alpha_n)^\top \end{pmatrix} \qquad \Phi' = \begin{pmatrix} \phi(s_1{}',\pi(s_1'))^\top \\ \vdots \\ \phi(s',\pi(s'))^\top \\ \vdots \\ \phi(s_n{}',\pi(s_n'))^\top \end{pmatrix}$$

Επομένως η συνάρτηση αξίας κατάστασης-ενέργειας και η συνάρτηση ανταμοιβής μπορουν να γραφούν σε αλγεβρική μορφή ως:

$$Q^\pi = \Phi w \tag{4.2}$$

$$R = \Phi u \tag{4.3}$$

## Εύρεση της πολιτικής (ευθύ πρόβλημα)

Από την εξίσωση *Bellman* για την συνάρτηση αξίας κατάστασης-ενέργειας προκύπτει:

$$Q = R + \gamma Q' \text{ αντικαθιστώντας τις 4.2 και 4.3} \tag{4.4}$$

$$\Phi w = \Phi u + \gamma \Phi' w \tag{4.5}$$

Ορίζεται έτσι το πρόβλημα εκτίμησης των παραμέτρων ελαχίστων τετραγώνων:

$$\min_{w,u} \frac{1}{2} ||\Phi w - (\Phi u + \gamma \Phi' w)||^2 \tag{4.6}$$

που είναι ένα πρόβλημα ελαχιστοποίησης με δύο αγνώστους παραμέτρους. Για την επίλυση του χρησιμοποιείται το εξής αλγοριθμικό σχήμα. Αρχικά θεωρούμε σταθερό το $u$ οπότε η μόνη άγνωστη παράμετρος είναι το $w$, δηλαδή είναι ένα πρόβλημα εκτίμησης πολιτικής. Τότε η λύση του προβλήματος για το $w$ προκύπτει από τον εκτιμητή ελαχίστων τετραγώνων θέτωντας την παράγωγο ως προς $w$ ίση με το μηδέν. Οπότε προκύπτει:

$$\widehat{w} = {A_w}^{-1} b_w \tag{4.7}$$

όπου

$$A_w = \Phi(\Phi - \gamma \Phi')^\top \tag{4.8}$$

$$b_w = \Phi R \tag{4.9}$$

$$\tag{4.10}$$

με $A_w$ έναν $(k \times k)$ πίνακα και $b_w$ ένα διάνυσμα $(k \times 1)$.

## Εύρεση της ανταμοιβής (αντίστροφο πρόβλημα)

Έχοντας εκτιμήσει το $w$ το επόμενο στάδιο περιλαμβάνει την εκτίμηση των παραμέτρων $u$. Διατηρώντας σταθερό το $\widehat{w}$ το πρόβλημα μετατρέπεται σε πρόβλημα εκτίμησης ανταμοιβής. Υπολογίζουμε από το σύνολο παραδειγμάτων $D$ για κάθε κατάσταση $s_i, {s_i}'$ τη συνάρτηση αξίας κατάστασης ενέργειας $Q$ και $Q'$ αντίστοιχα. Επομένως,

$$\widehat{Q} = \widehat{Q} - \gamma \widehat{Q'} \tag{4.11}$$

με $\widehat{Q} = \Phi \widehat{w}$ και $\widehat{Q'} = \Phi' \widehat{w}$. Το πρόβλημα βελτιστοποίησης της εξίσωσης 4.5 μετρέπεται σε:

$$\min_{u} \frac{1}{2} ||\Phi u - \Delta \widehat{Q})||^2 \tag{4.12}$$

Η λύση του προκύπτει από τον εκτιμητή ελαχίστων τετραγώνων θέτωντας την παράγωγο ως προς $u$ ίση με το μηδέν. Οπότε :

$$\widehat{u} = {A_u}^{-1} b_u \tag{4.13}$$

όπου

$$A_u = \Phi(\Phi)^\top \tag{4.14}$$

$$b_u = \Delta\widehat{Q}\Phi \tag{4.15}$$

με $A_u$ έναν $(k \times k)$ πίνακα και $b_u$ ένα διάνυσμα $(k \times 1)$. Το παραπάνω σχήμα είναι επαναληπτικό με τις εξισώσεις 4.6 και 4.12 να εναλλάσονται σε κάθε επεισόδιο μέχρι να ικανοποιείται ένα κριτήριο τερματισμού. Αφού βρεθεί η πολιτική με βάση

Τα δείγματα $D$ που συλλέχθηκαν χρησιμοποιούνται σε κάθε επανάληψη του αλγορίθμου για την εκτίμηση των παραμέτρων $w$ και $u$. Έπειτα με βάση την τρέχουσα πολιτική δημιουργεί ένα νέο σύνολο δειγμάτων $D'$ και χρησιμοποιώντας το εκτιμά εκ νέου τις παραμέτρους $w, u$. Αυτή η διαδικασία επαναλαμβάνεται μέχρις ότου να ικανοποιείται ένα κριτήριο σύγκλισης.

### 4.4.2 Online Μάθηση

Ο αλγόριθμος που παρουσιάστηκε παραπάνω είναι ένας *off-line* αλγόριθμος μιας και δέχεται ως είσοδο ένα σύνολο απο παραδείγματα έτσι ώστε να ανακαλύψει την βελτιστη πολιτική και την συνάρτηση ανταμοιβής. Έχουμε δημιουργήσει μια προσεγγίση ώστε να μετατρέψουμε αυτόν τον αλγόριθμο σε *on-line*

Κατά τη διαδικασία εκτίμησης της πολιτικής οι πίνακες $A_w, A_u$ και τα διάνυσματα $b_w, b_u$ μπορούν να γραφούν και με την ακόλουθη μορφή: :

$$A_w = \sum_{i=1}^{n} \phi(s_i, \alpha_i)(\phi(s_i, \alpha_i) - \gamma\phi(s_i', \pi(s_i')))^\top \tag{4.16}$$

$$b_w = \sum_{i=1}^{n} \phi(s_i, \alpha_i) r_i \tag{4.17}$$

$$A_u = \sum_{i=1}^{n} \phi(s_i, \alpha_i)(\phi(s_i, \alpha_i))^\top \tag{4.18}$$

$$b_u = \sum_{i=1}^{n} w_i(\phi(s_i, \alpha_i) - \gamma\phi(s_i', \alpha_i'))^\top \phi(s_i, \alpha_i) \tag{4.19}$$

Εκμεταλευόμενοι τις σχέσεις (4.15),(4.16),(4.17) και (4.18) οι $A_w, A_u$ και $b_w, b_u$ μπορούν να δημιουργούνται αυξητικά σε κάθε βήμα του αλγορίθμου για την εκτί-

μηση των παραμέτρων $w, u$. Για ένα νέο δείγμα τη χρονική στιγμή $t$ $(s_t, \alpha_t, r_t, s_t{}')$ υπολογίζονται ως :

$$\begin{aligned} A_w &= A_w + \phi(s_t, \alpha_t)(\phi(s_t, \alpha_t) - \gamma\phi(s_t{}', \alpha_t{}'))^\top, \\ b_w &= b_w + \phi(s_t, \alpha_t)\ r_t, \end{aligned} \tag{4.20}$$

όπου $s$ η κατάσταση στην οποία ήδη βρισκόμαστε και $s'$ η κατάσταση στην οποία μετάβαίνουμε έπειτα από την εκτέλεση της ενέργειας $\alpha$. Η απολαβή που λαμβάνουμε συμβολίζεται ως $r_t$. Αν η νέα κατάσταση στην οποία βρεθούμε είναι τερματικη, τότε υπάρχει μια διαφοροποίηση στον κανόνα ενημέρωσης του πίνακα και η μορφή του είναι:

$$A_w \ = \ A_w + \phi(s_t, \alpha_t)(\phi(s_t, \alpha_t))^\top \tag{4.21}$$

Αφού οι πίνακες $A_w$ και $b_w$ έχουν υπολογιστεί, χρησιμοποιούνται για να υπολογιστούν τα βάρη $w$ ώστε να παραχθεί μια πολιτική για την εκτίμηση της συνάρτησης αξίας κατάστασης-ενεργειας $Q$. Αυτό επιτυχγάνεται υπολογίζοντας το διανυσμα $w$ ως.

$$w = A_w^{-1} b_w. \tag{4.22}$$

Όπως αναφέραμε αρχικά ταυτόχρονα με την εύρεση της πολιτικής στοχεύουμε και στην εύρεση των ανταμοιβών. Για την εύρεση της συνάρτησης ανταμοιβών χρησιμοποιούμε παρόμοια λογική με την εύρεση των $w$. Οι κανόνες ενημέρωσης που χρησιμοποιούμε είναι:

$$\begin{aligned} A_u &= A_u + \phi(s_t, \alpha_t)(\phi(s_t, \alpha_t))^\top \\ b_u &= b_u + Q\ \phi(s_t, \alpha_t), \end{aligned} \tag{4.23}$$

Το $\Delta Q$ υπολογίζεται ως εξης:
Αν η νέα κατάσταση είναι τερματική τότε

$$\Delta Q = \phi(s_t, \alpha_t)^\top w \tag{4.24}$$

αλλιώς σε κάθε άλλη περιπτωση

$$\Delta Q = (\phi(s_t, \alpha_t) - \gamma\phi(s_t{}', \alpha_t{}'))^\top w \tag{4.25}$$

Έπειτα απο την ενημέρωση των $_u$ και $b_u$, υπολογίζουμε το διάνυσμα $u$

$$u = A_u^{-1} b_u \tag{4.26}$$

το οποίο περιλαμβάνει τα βάρη για την εύρεση της συνάρτησης ανταμοιβής με τον εξής κανονα:

$$R(s_t, \alpha_t) = \phi(s_t, \alpha_t)^\top u \tag{4.27}$$

Η παραπάνω διαδικασία επαναλμβάνεαι για κάθε νέο δείγμα που παρατηρεί ο πράκτορας εως ότου να ικανοποιείται ένα κρίτηριο σύγκλισης των $w$ και $u$.

### 4.4.3 Το αλγοριθμικό σχήμα της προτεινόμενης μεθοδολογίας

Ο αλγόριθμος που ακολουθεί σε μορφή ψευδοκώδικα περιγράφει το προτεινόμενο αλγοριθμικό σχήμα.

**Αλγόριθμος 4.1** Η online προτεινόμενη μέθοδος

$A_w \leftarrow 0_{k\times k}$

$A_u \leftarrow 0_{k\times k}$

$b_w \leftarrow 0_k$

$b_w \leftarrow 0_k$

$u \leftarrow random_k$

**repeat** (για κάθε επεισόδιο)

Αρχικοποίηση της κατάστασης $s_t$, με t=0

**repeat** (για κάθε βήμα του επεισοδίου)

Επιλογή της ενέργειας $\alpha_t$ στην καταστση $s_t$

Εκτέλεση της ενέργειας $\alpha_t$

Υπολογισμός της ανταμοιβή $r = \phi(s_t, \alpha_t)u$

Μετάβαση στην νέα κατάσταση $s_{t+1}$

**if** $s_{t+1}$ **μη τερματική κατάσταση**

$$\begin{aligned} A_w &= A_w + \phi(s_t,\alpha_t)(\phi(s_t,\alpha_t) - \gamma\phi(s_{t+1},\alpha_{t+1}))^\top \\ b_w &= b_w + \phi(s_t,\alpha_t)\ r \\ A_u &= A_u + \phi(s_t,\alpha_t)(\phi(s_t,\alpha_t))^\top \\ \Delta Q &= (\phi(s_t,\alpha_t) - \gamma\phi(s_{t+1},\alpha_{t+1})w)^\top \\ b_u &= b_u + \Delta Q\ \phi(s_t,\alpha_t) \end{aligned}$$

**else**

$$\begin{aligned} A_w &= A_w + \phi(s_t,\alpha_t)(\phi(s_t,\alpha_t))^\top \\ b_w &= b_w + \phi(s_t,\alpha_t)\ r \\ A_u &= A_u + \phi(s_t,\alpha_t)(\phi(s_t,\alpha_t))^\top \\ \Delta Q &= \phi(s_t,\alpha_t)^\top w \\ b_u &= b_u + \Delta Q\ \phi(s_t,\alpha_t) \end{aligned}$$

**end**

$$\begin{aligned} w &= A_w^{-1} b_w \\ u &= A_u^{-1} b_u \\ t &= t+1 \end{aligned}$$

**until** (τέλος του επεισοδίου)

**until** (σύγκλιση)

# Κεφαλαιο 5

# Πειραματικη Αξιολογηση



## 5.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε τα αποτελέσματα της πειραματικής αξιολόγησης της προτεινόμενης μεθόδου αλλά και των μεθόδων ενισχυτικής μάθησης που περιγράψαμε στην παρούσα διατριβή. Αρχικά, θα περιγράψουμε την πειραματική διαδικασία καθώς και το πρόβλημα που κληθήκαμε να επιλύσουμε,τα πειραματικά περιβάλλοντα που χρησιμοποιήσαμε και τέλος τα αποτελέσματα που λάβαμε από την αξιολόγησή των μεθόδων.

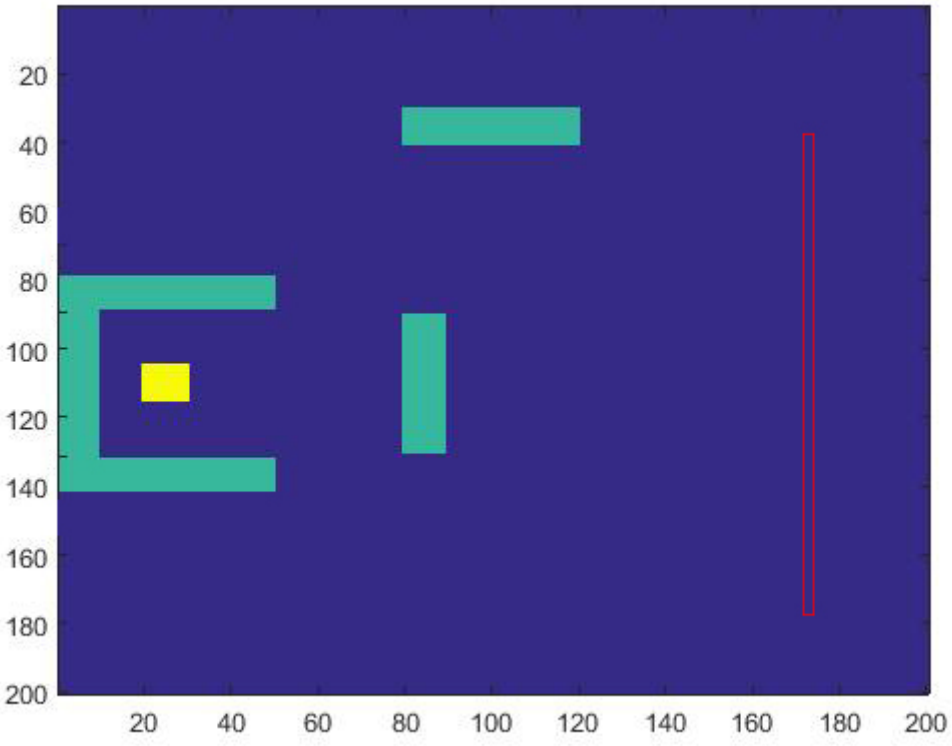
## 5.2 Πειραματική Διαδικασία

Στην παρούσα διατριβή μελετήσαμε το πρόβλημα της πλοήγησης της θαλάσσιας ρομποτικής πλατφόρμας προς μια επιθυμητή θέση. Κατά τη διάρκεια της πλοήγησης,

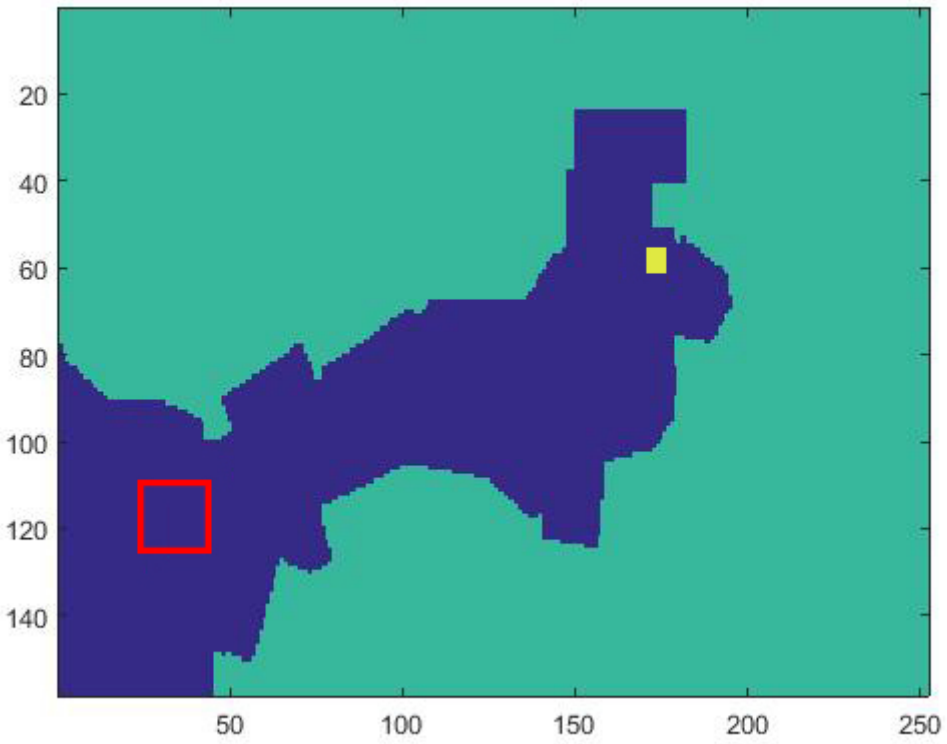
το περιβάλλον μεταβάλλεται δυναμικά. Περιελαμβάνει το κινηματικό και δυναμικό μοντέλο της θαλάσσιας ρομποτικής πλατφόρμας αλλά και τις περιβαλλοντικές διαταραχές που επιδρούν πάνω της κάθε χρονική στιγμή $\delta t$. Το $\delta t$ όριζεται ως το βήμα ολοκλήρωσης που χρησιμοποιούμε για τον υπολογισμό της θέσης αλλά και της ταχύτητας της πλατφόρμας και ισούται με $0.2$. Οι περιβαλλοντικές διαταραχές είναι ο άνεμος, τα κύματα της θάλασσας καθώς και τα θαλάσσια ρεύματα. Το περιβάλλον προσομοίωσης υλοποιήθηκε με την χρήση του προγραμματιστικού περιβάλλοντος της MATLAB. Εξαιτίας της μεγάλης μάζας της ρομποτικής πλατφόρμας κάναμε κάποιες συμβάσεις. Η μετάβασή της σε μια νέα κατάσταση δεν μπορεί να πραγματοποιηθεί σε μία χρονική στιγμή $\delta t$. Για το λόγο αυτό διατηρούμε για κάποιο χρονικό διάστημα $(100\delta t)$ την ίδια ενέργεια προς εκτέλεση από την ρομποτική πλατφόρμα έτσι ώστε η νέα θέση στην οποία θα μεταβεί να διαφέρει αρκετά από αυτή που ήδη βρισκόταν. Βέβαια το χρονικό διάστημα αυτό εξαρτάται από την συνολική δύναμη που παράγουν τα τζετ-νερού. Επίσης, ο υπολογισμός του διανύσματος των παραμέτρων $w$ για την εύρεση της πολιτικής αλλά και της συνάρτησης ανταμοιβής επιλέξαμε να υπολογίζονται αφού πρώτα ο πράκτορας-ρομποτική πλατφόρμα πραγματοποιήσει έναν ορισμένο αριθμό βημάτων. Ο αριθμός αυτών των βημάτων επιλέχθηκε να είναι 5 έπειτα από πειραματική προσέγγιση έτσι ώστε οι πίνακες για την εύρεση τόσο της πολιτικής όσο και της συνάρτησης ανταμοιβής να συλλέγουν αρκετή πληροφορία. Κατά τη διάρκεια της διαδικασίας μάθησης ένα καινούργιο επεισόδιο ξεκινά είτε όταν η πλατφόρμα προσκρούσει σε κάποιο εμπόδιο, είτε όταν εντοπίσει τον στόχο, είτε βρεθεί εκτός των ορίων του κάθε χάρτη. Επίσης κριτήριο για την έναρξη ενός νέου επεισοδίου αποτελεί και η άσκοπη περιηγησή, οριοθετώντας τον μέγιστο επιτρεπτό αριθμό βήμάτων σε 1000.

## 5.3 Πειραματικά Περιβάλλοντα

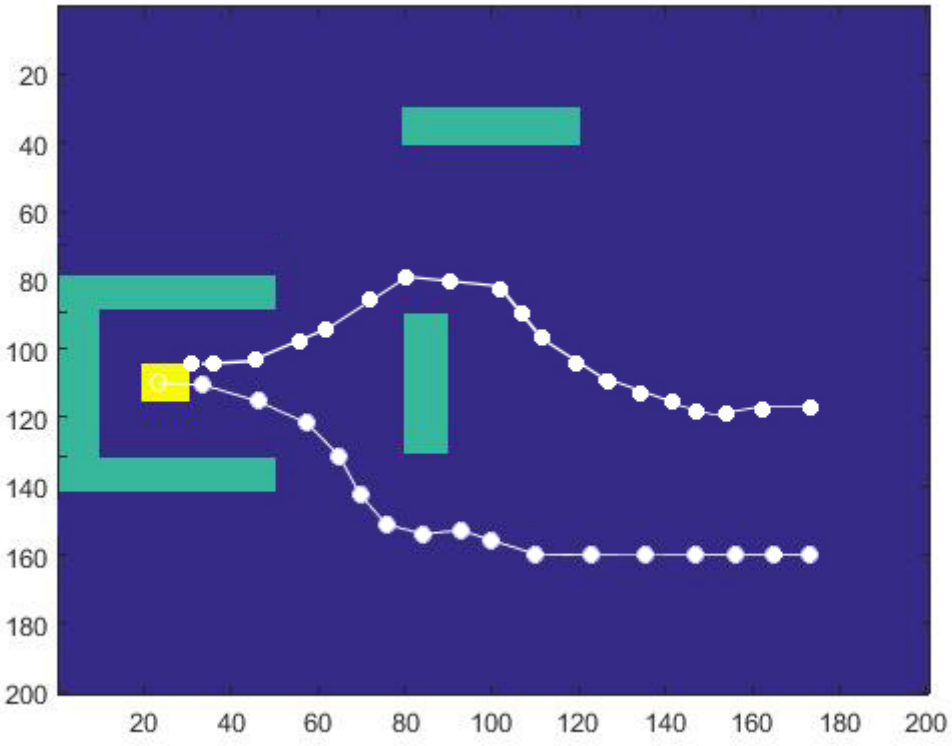
Για την πειραματική αξιολόγηση των μεθόδων της παρούσας διατριβής χρησιμοποιήθηκαν δύο περιβάλλοντα. Το πρώτο αποτελείται από έναν τεχνητό χάρτη, (Σχήμα 5.1) που κατασκευάσαμε ενώ το δεύτερο παρουσιάζει ένα πραγματικό περιβάλλον που απεικονίζει το λιμάνι του Πειραιά, (Σχήμα 5.2). Η εικόνα του λήφθηκε από τους χάρτες της *Google* και προσαρμόστηκε στο πρόβλημά μας έπειτα από τεχνικές επεξεργασίας εικόνας στην οποία υποβλήθηκε.



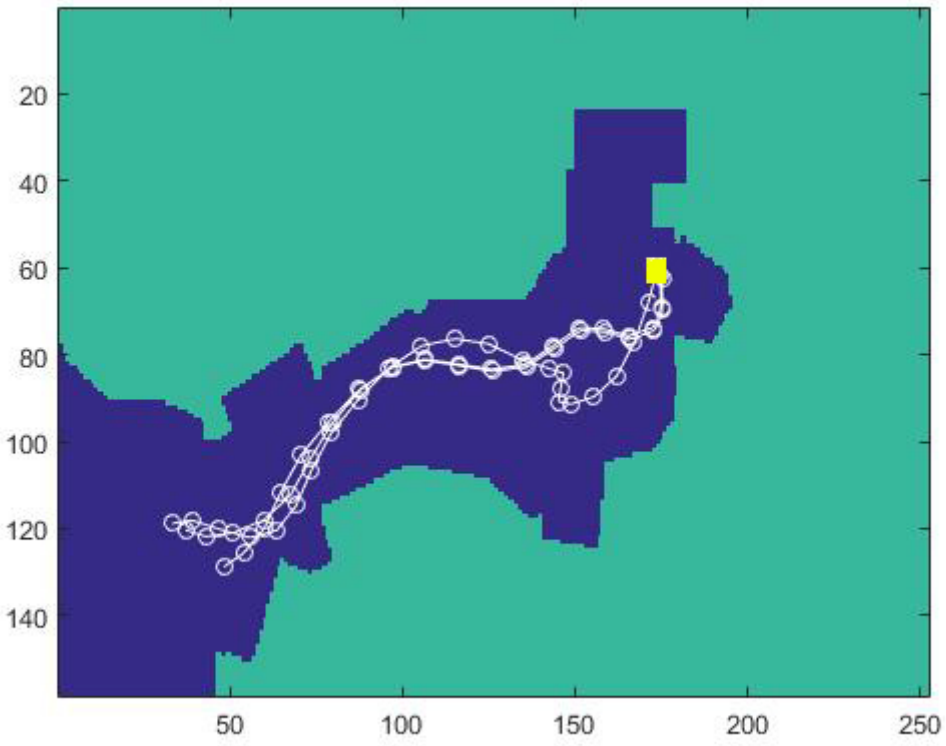

Σχήμα 5.1: Το τεχνητό πειραματικό περιβάλλον




Σχήμα 5.2: Ο χάρτης του λιμανιού του Πειραιά




Σχήμα 5.3: Χαρακτηριστικές τροχιές στον τεχνητό χάρτη




Σχήμα 5.4: Χαρακτηριστικές τροχιές στον χάρτη του Πειραιά

Ο τεχνητός χάρτης, (Σχήμα 5.1), έχει διαστάσεις $200 \times 200$ $pixels$ ενώ ο χάρτης του Πειραιά, (Σχήμα 5.2) $252 \times 158$ $pixels$. Στους δύο χάρτες με μπλε χρώμα συμβολίζεται ο χώρος που μπορεί να κινηθεί η θαλάσσια ρομποτική κατασκευή. Με πράσινο, στον τεχνητό χάρτη, αναπαρίστανται τα εμπόδια που καλείται να αποφύγει έτσι ώστε να μην υπάρχει σύγκρουση, ενώ με το ίδιο χρώμα στον χάρτη του Πειραιά οριοθετούμε το χώρο του λιμανιού. Τέλος, με κίτρινο χρώμα συμβολίζεται ο στόχο προς επίτευξη.

Σκοπός της διατριβής μας είναι η πλοήγηση του θαλάσσιου ρομποτικού συστήματος σε καθέναν από τους δύο χάρτες έτσι ώστε να εντοπίσει το συντομότερο μονοπάτι από οποιοδήποτε σημείο βρίσκεται εντός του κόκκινου πλαισίου, να αποφεύγει επιτυχώς τα εμπόδια που θα συναντήσει στην πορεία του και να φθάνει στο εκάστοτε στόχο. Η πλοήγησή του όμως επηρεάζεται κι από τις διάφορες περιβαλλοντικές διαταραχές που περιγράφαψε στο κεφάλαιο 3. Η ανταμοιβή που λαμβάνει ο πράκτορας σε κάθε βήμα είναι -1. Στην περιπτωση που φθάνει στην τερματική κατάσταση-στόχο η ανταμοιβή του είναι 100, ενώ στην περίπτωση που βρίσκεται σε απόσταση μικρότερη των 2 *pixels* από κάποιο εμπόδιο -100. Τέλος, τερματική κατάσταση θεωρούμε και την εκτός ορίων θέση του κάθε χάρτη στην οποία πιθανώς να βρεθεί η ρομποτική πλατφόρμα. Η ανταμοιβή που λαμβάνει σε αυτή την περίπτωση ο πράκτορας είναι -500. Αυτές τις ανταμοιβές τις χρησιμοποιήσαμε στην εφαρμογή των αλγορίθμων *LSPI* και *Q-Learning*. Στην περίπτωση που ο πράκτορας φθάσει σε μια τερματική κατάσταση, μια από αυτές που αναφέραμε παραπάνω, το επεισόδιο τερματίζει και ξεκινά ένα καινούργιο από μια τυχαία αρχική θέση εντός των κόκκινων πλαισίων. Στόχος του πράκτορα είναι να φθάσει στην κατάσταση-στόχο με τα λιγότερα βήματα. Κάποιες χαρακτηριστικές τροχιές της πλατφόρμας στους δύο χάρτες παρουσιάζονται στα Σχήματα 5.3 και 5.4

## 5.4 Η επίδραση του πλήθους χαρακτηριστικών και ενεργειών στην επίδοση της μεθόδου

### 5.4.1 Επίδραση του πλήθους των χαρακτηριστικών

Ένα σημαντικό ζήτημα της πειραματικής διαδικασίας αποτέλεσε η επιλογή του πλήθους των χαρακτηριστικών που θα χρησιμοποιούνταν. Για το λόγο αυτό αξιολογήσαμε την επίδραση τους στη σύγκλιση της μεθόδου. Η ακόλουθη γραφική πα-

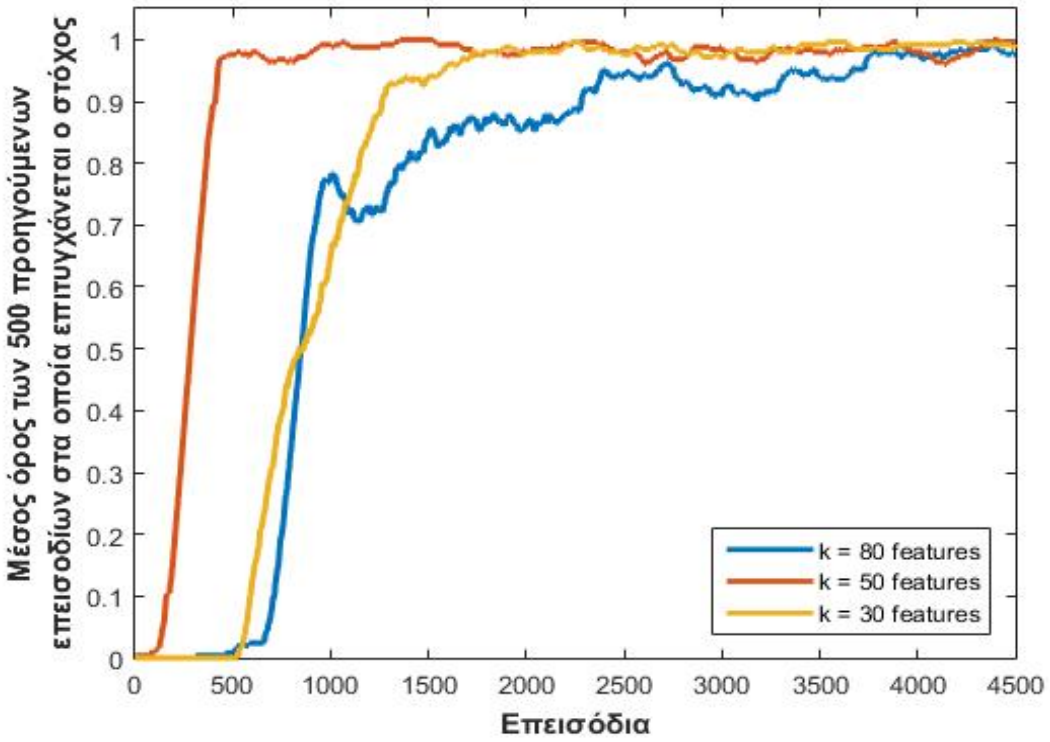
ράσταση, Σχήμα 5.5, αναπαριστά των μέσο όρο των 500 προηγούμενων επεισοδίων στα οποία επιτυγχάνεται ο στόχος στον χάρτη Πειραιά χωρίς την επίδραση των περιβαλλοντικών διαταραχών. Εφαρμόστηκε το προτεινόμενο αλγοριθμικό σχήμα κι αξιολογήθηκε με διαφορετικό πλήθος χαρακτηριστικών. Όπως παρατηρείται με 50 χαρακτηριστικα η διαδικασία μάθησης επιτυχγάνεται ταχύτερα σε σχέση με με την επιλογή λιγότερων (30) ή περισσότερων (80) χαρακτηριστικών. Φυσικά το πλήθος τους αποτελεί κρίσιμο παράγωντα. Λιγότερα χαρακτηριστικά μπορεί να οδηγήσουν σε *underfitting* ενώ η επιλογή περισσότερων σε *overfitting*. Έπειτα από αυτή την μελέτη η συνέχεια της πειραματικής διαδικασίας πραγματοποιείται με 50 χαρακτηριστικά.

### 5.4.2 Επίδραση του πλήθους των ενεργειών

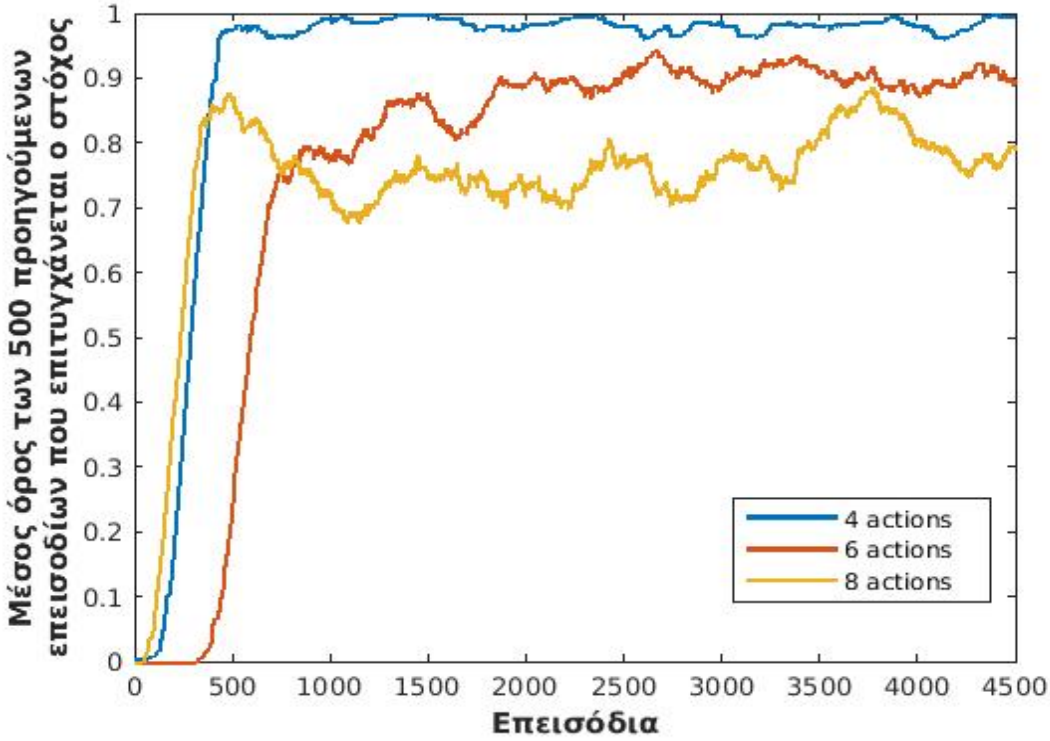
Εκτός από την επίδραση του πλήθους του χαρακτηριστικών, μελετήσαμε και την επίδραση του πλήθους των ενεργειών που είναι διαθέσιμενες προς εκτέλεση από τη ρομποτική πλατφόρμα κάθε χρονική στιγμή. Η εφαρμογή διαφορετικού εύρους ενεργειών πραγματοποιήθηκε στο προτεινόμενο αλγοριθμικό σχήμα που περιγράψαμε στο κεφάλαιο 4 χωρίς τη χρήση περιβαλλοντικών διαταραχών. Ο αριθμός των συναρτήσεων βάσης που χρησιμοποιήσαμε ήταν 50 και το πλήθος των διαφορετικών ενεργειών-γωνιών περιγράφονται στον παρακάτω πίνακα.

| **Πλήθος Ενεργειών** | **Κατεύθυνση ( σε μοίρες )** |
|---|---|
| 4 | 0, 90, 180, 270 |
| 6 | 0, 60, 120, 180, 240, 300 |
| 8 | 0, 45, 90, 135, 180, 225, 270, 315 |

Όπως γίνεται αντιληπτό παρατηρώντας την παρακάτω γραφική παράσταση, Σχήμα 5.6, με την επιλογή τεσσάρων γωνιών επιτυγχάνεται η υψηλότερη απόδοση του αλγορίθμου. Η καλύτερη συμπεριφορά με μικρό πλήθος γωνιών οφείλεται στο μικρότερο χώρο καταστάσεων που καλείται να μάθει ο πράκτορας . Όσο αυξάνει το πλήθος αυτό αυξάνουν οι καταστάσεις και ο πράκτορας αργεί να ανακαλύψει μια βέλτιστη πολιτική.




Σχήμα 5.5: Μελέτη της επίδρασης του πλήθους των χαρακτηριστικών.




Σχήμα 5.6: Μελέτη της επίδρασης του πλήθους των ενεργειών.

## 5.5 Συγκριτικά αποτελέσματα

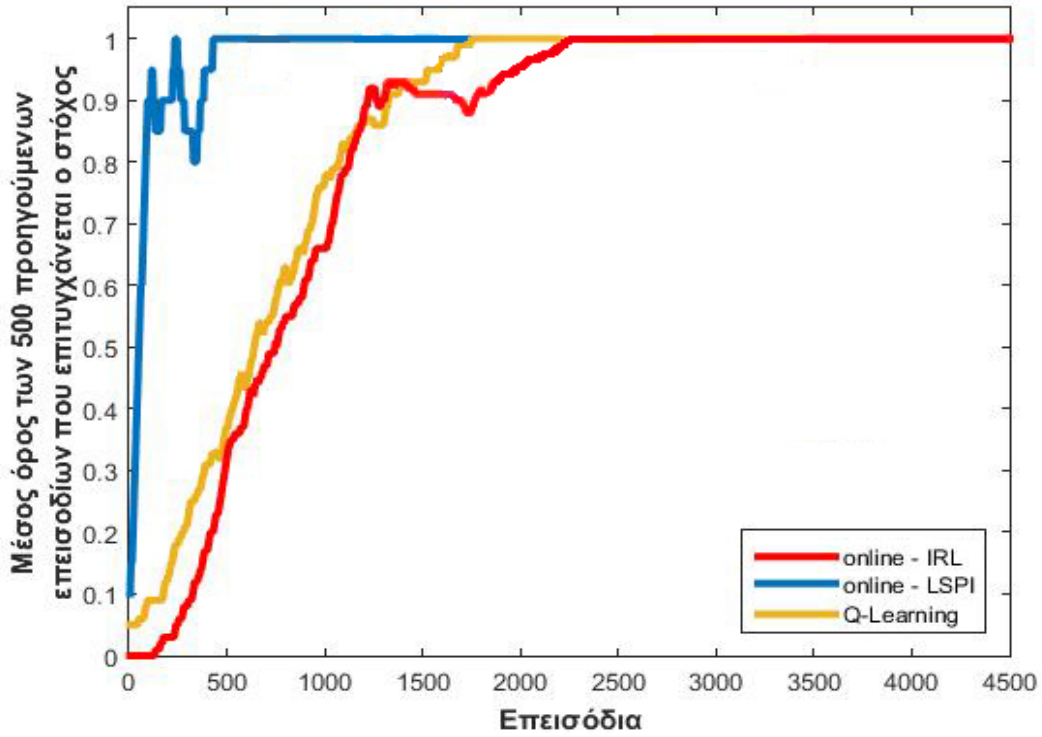
### 5.5.1 Περιβάλλον Ι

Αρχικά, για την επίλυση του προβλήματος πλοήγησης, υλοποιήσαμε τον αλγόριθμο χρονικών διαφορών *Q-learning*. Την διακριτοποίση του χώρου καταστάσεων την δημιουργήσαμε με τον ακόλουθο τρόπο. Στον τεχνητό χάρτη στο χώρο που μπορεί να κινηθεί η πλατφόρμα εφαρμόσαμε τον αλγόριθμο *k-means* με πλήθος κέντρων τον αριθμό 50. Έπειτα τις συντενταγμένς των κέντρων που επέστρεφε η εφαρμογή του αλγορίθμου *k-means* τα θεωρούσαμε ως διακριτές καταστάσεις δημιουργώντας με αυτο τον τρόπο το χώρο καταστάσεων $S$. Ως ενέργειες, επιλεξαμε τις γωνίες $0, 90, 180, 270$. Οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου *Q-learning* βρέθηκαν πως είναι οι εξής : ρυθμός μάθησης $\alpha = 0.1$, ρυθμός έκπτωσης $\gamma = 0.95$ και πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.3$. Το μέτρο της δυναμης που αρκούν συνολικά τα τζετ-νερού ήταν σταθερό στην τιμή $20000\ N$.

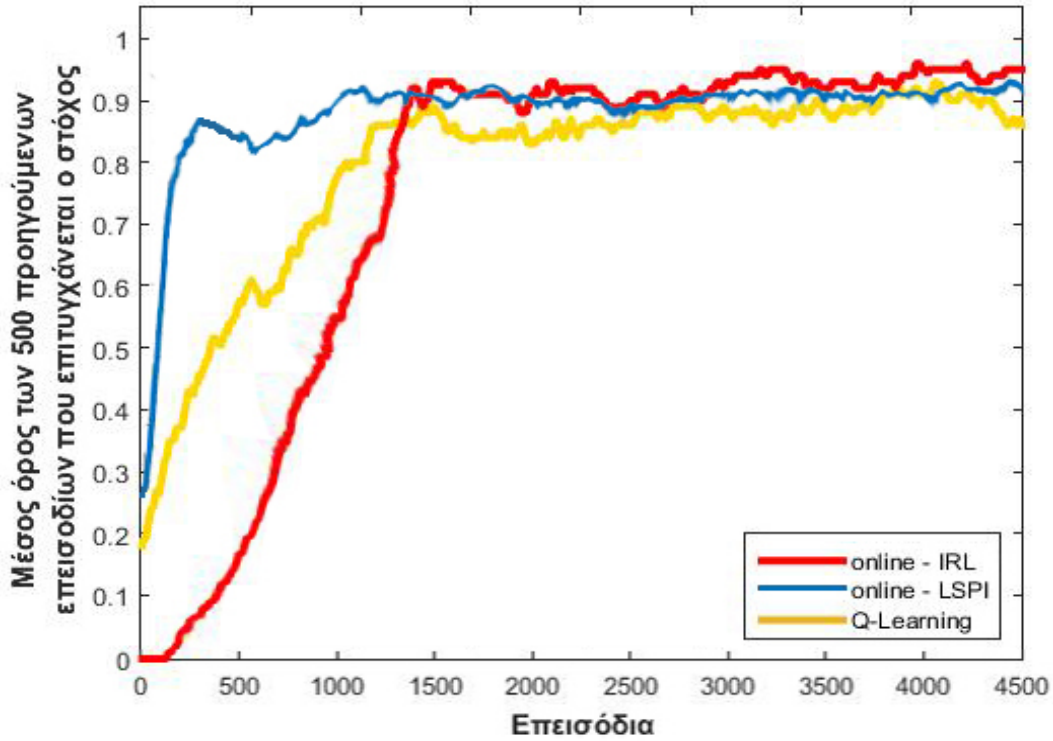
Επίσης στο ίδιο περιβάλλον εφαρμόστηκε κι ο αλγόριθμος online LSPI. Τα χαρακτηριστικά του αλγορίθμου *LSPI* παράχθηκαν με την χρήση 50 ακτινικών συναρτήσεων βάσης. Τα κέντρα των ακτινικών συναρτήσεων δημιουργήθηκαν με την χρήση του k-means όπως περιγράψαμε παραπάνω, ενώ η διακύμανση $\sigma^2$ επιλέχθηκε να είναι σταθερή στην τιμή 0.01. Η εύρεση της βέλτιστης τιμή της έγινε με μια πειραματική προσέγγιση. Οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου online-LSPI βρέθηκαν πως είναι οι εξής : ρυθμός έκπτωσης $\gamma = 0.95$, πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.3$. Το πλήθος ενεργειών παρέμεινε το ίδιο.

Τέλος, το προτεινόμενο αλγοριθμικό σχήμά εφαρμόστηκε με βέλτιστες παραμέτρους όμοιες με αυτές του αλγορίθμου *online-LSPI*. Κατά την εφαρμογή των μεθόδων *online-LSPI* και *Q-Learning* η τιμή της ανταμοιβής ήταν προκαθορισμένη. Σε αυτο, ο πράκτορας καλείται να ανακαλύψει εκτός από την βέλτιση πολιτικη και την τιμή των ανταμοιβών. Κατά την εκτέλεση του συγκεκριμένου αλγορίθμου επιτρέπουμε σταθερές ανταμοιβές στον πράκτορα για τα πρώτα 100 επεισόδια. Έπειτα αναλαμβάνει να ανακαλύψει μόνος του τις τιμές των ανταμοιβών.

Στις ακόλουθες γραφικές παραστάσεις, Σχήματα 5.7(a),(b), παρουσιάζονται συγκριτικά αποτελέσματα μεταξύ των τριών μεθόδων χωρίς περιβαλλοντικές διαταραχές αλλά και με τη χρήση τους. Μπορούμε εύκολα να παρατηρήσουμε ότι το προτεινόμενο αλγοριθμικό σχήμα παρουσιάζει εξίσου καλή συμπεριφορά σύγκλισης με τους άλλους δύο αλγορίθμους.




(a) Ποσοστό επιτυχίας εύρεσης στόχου των 500 προηγούμενων επεισοδίων στον τεχνητό χάρτη χωρίς περιβαλλοντικές διαταραχές



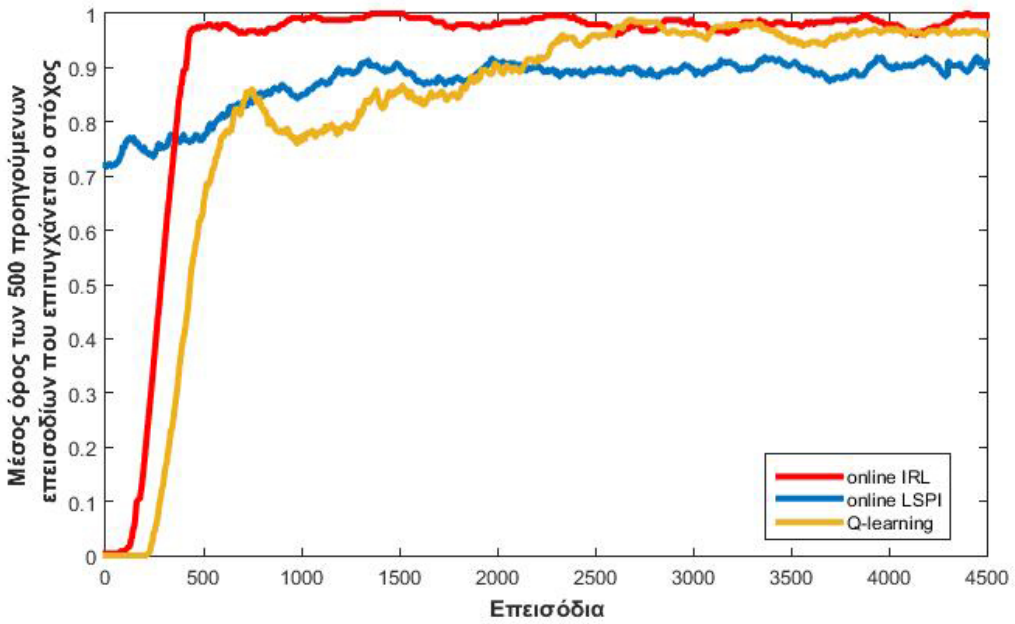

(b) Ποσοστό επιτυχίας εύρεσης στόχου των 500 προηγούμενων επεισοδίων στον τεχνητό χάρτη με περιβαλλοντικές διαταραχές

Σχήμα 5.7: Συγκριτικά αποτελέσματα στον τεχνητό χάρτη

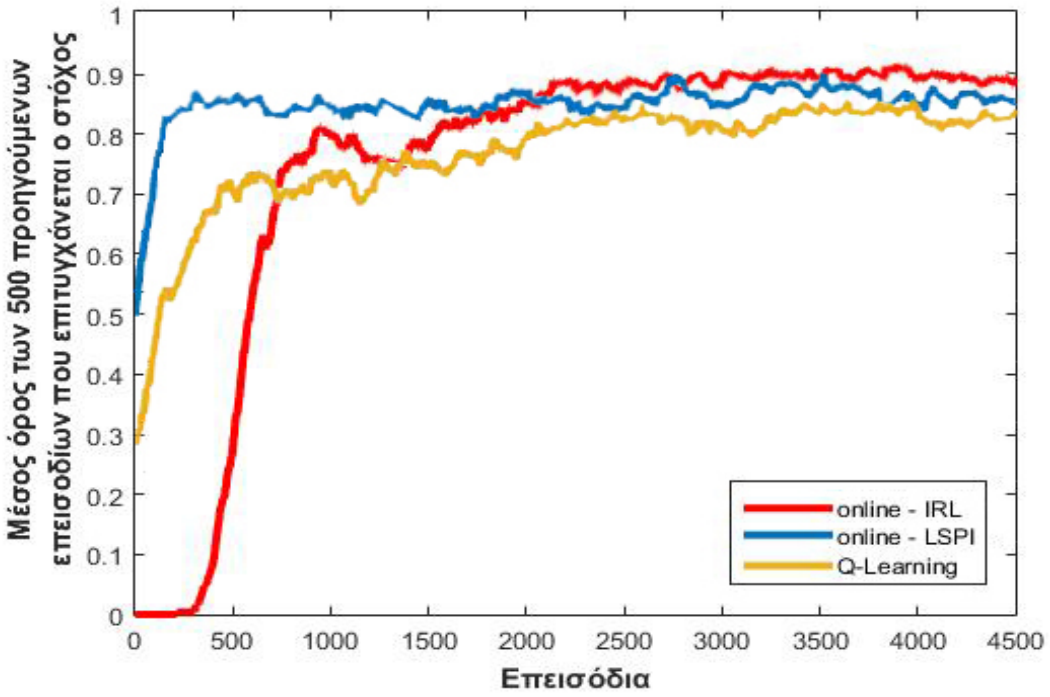
### 5.5.2 Περιβάλλον ΙΙ

Όπως και στο τεχνητό χάρτη, έτσι και στο χάρτη του Πειραιά εφραρμόσαμε τις τρεις μεθόδους που αναφέραμε παραπάνω. Στο σημείο αυτό οφείλουμε να αναφέρουμε ότι ο αλγόριθμος Q-Learning με χώρο καταστάσεων 50 διακριτά σημεία δεν κατάφερνε να προσεγγίσει τον στόχο είτε με την χρήση διαταραχών είτε όχι. Για να παρουσιάσει μια αξιόλογη συμπεριφορά χρειαζόταν πολύ περισσότερα σημεία όπως τελικα αποδειχθηκε έπειτα απο πειραματική προσέγγιση. Οι βέλτιστες τιμές των παραμέτρων παραμένουν ίδιες όπως και στον τεχνητό χάρτη.

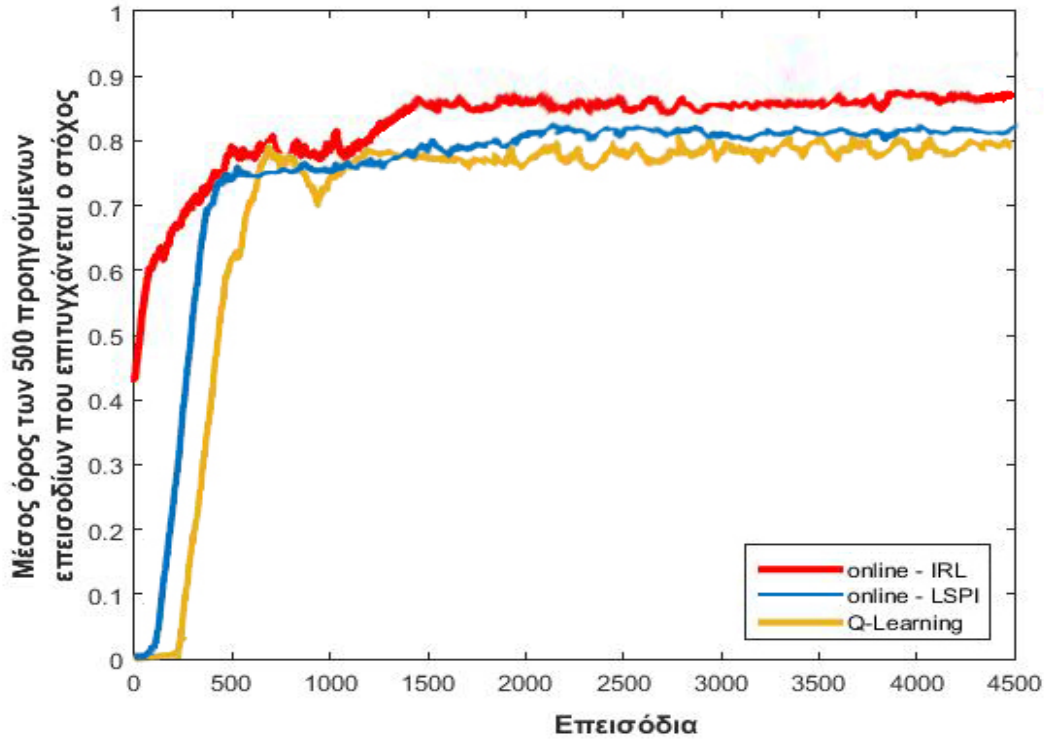
Στη συνέχεια παρουσιάζονται τα συγκριτικά αποτελέσμα μεταξύ των αλγορίθμων *online-LSPI*, της προτεινόμενης μεθόδου και του *Q-Learning*, Σχήματα 5.8, 5.9 και 5.10. Όπως μπορεί να γίνει έυκολα αντιληπτό παρατηρώντας τις δύο γραφικές παραστάσεις το προτεινόμενο αλγοριθμικό σχήμα χωρίς την επίδραση περιβαλλοντικών διαταραχών επιτυχγάνει πολύ καλύτερη απόδοση από τον *online LSPI*, ενώ πολύ καλή συμπεριφορά παρουσιάζει και με την χρήση περιβαλλοντικών διαταραχών με τις οποίες η πλοήγηση δυσκολεύει αρκετά. Όπως παρατηρείται η προτεινόμενη μέθοδος παρουσιάζει καλύτερη συμπεριφορά Τέλος, στο Σχήμα 5.11 απεικονίζονται οι τιμές της βέλτισης πολιτικής αλλά και της συνάρτησης ανταμοιβής έπειτα από εφαρμογή του προτεινόμενου αλγορίθμου στον χάρτη του Πειραιά.



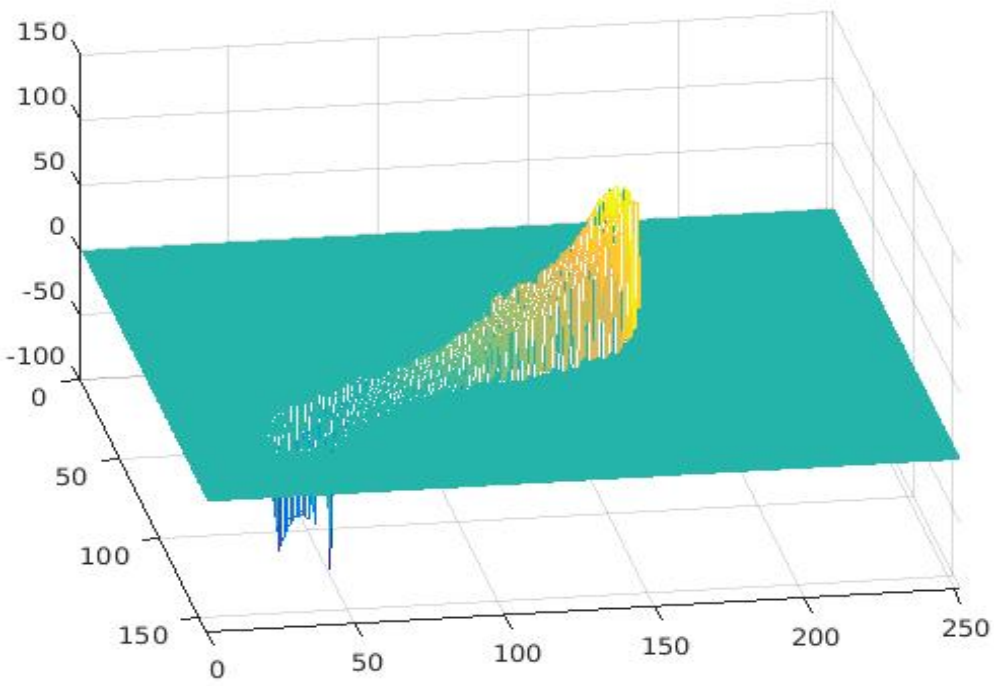

Σχήμα 5.8: Συγκριτικά αποτελέσματα στον χάρτη του Πειραιά χωρίς περιβαλλοντικές διαταραχες



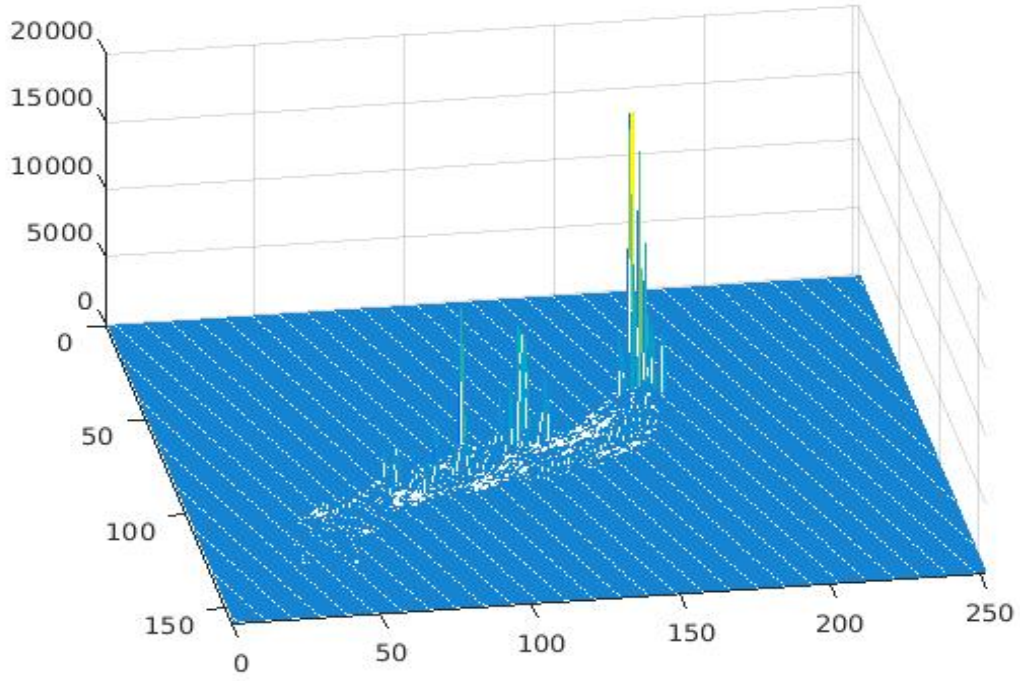

Σχήμα 5.9: Συγκριτικά αποτελέσματα στον χάρτη του Πειραιά με περιβαλλοντικές διαταραχες εξαιτίας των κυμάτων και των θαλάσσιων ρευμάτων καθώς και με την προσθήκη του θορύβου μετρήσεων GPS




Σχήμα 5.10: Συγκριτικά αποτελέσματα στον χάρτη του Πειραιά με περιβαλλοντικές διαταραχες




(a) Γραφική απεικόνιση της βέλτιστης πολιτικής του προτεινόμενου αλγορίθμου στο χάρτη του Πειραιά




(b) Γραφική απεικόνιση της βέλτιστης συνάρτησης ανταμοιβής του προτεινόμενου αλγορίθμου στο χάρτη του Πειραιά

Σχήμα 5.11: Πολιτική και Συνάρτηση Ανταμοιβής

# Κεφαλαιο 6

# Συμπερασματα

---

Στην παρούσα διατριβή επικεντρωθήκαμε στο πρόβλημα της αυτόνομης πλοήγησης της θαλάσσιας ρομποτικής πλατφόρμας Delta Berenike με χρήση μεθόδων ενισχυτικής μάθησης. Κατά τη διάρκεια της πλοήγησης, οι περιβαλλοντικές δυνάμεις που επιδρούν πάνω της δυσκολεύουν τη μετακίνησή της. Η ενισχυτική μάθηση σχετίζεται με τον τρόπο που ένας πράκτορας μπορεί να μάθει μια πολιτική μέσω της αλληλεπίδρασής του με το περιβάλλον για την επίτευξη ενός στόχου.

Αρχικά, μελετήσαμε το πλαίσιο της ενισχυτικής μάθησης, καθώς και βασικές μεθόδους που σχετίζονται με αυτή. Στη συνέχεια, εστιάσαμε στη συνάρτηση ανταμοιβής, η οποία καθορίζει τη συμπεριφορά ενός πράκτορα και ο ορισμός της αποτελεί ένα σημαντικό ζήτημα. Η εκτίμηση της συνάρτησης ανταμοιβής είναι ένα πρόβλημα το οποίο πραγματεύεται η αντίστροφη ενισχυτική μάθηση, η οποία δοθείσας μίας βέλτιστης πολιτικής αναζητά τις ανταμοιβές.

Η προσφορά της παρούσας διατριβής είναι η υλοποίηση μιας μεθόδου η οποία στοχεύει στην ταυτόχρονη εύρεση τόσο της βέλτιστης πολιτικής, όσο και των ανταμοιβών που λαμβάνει ο πράκτορας. Το κυριότερο πλεονέκτημά της είναι η γενικευτική ικανότητά της ως προς τη μορφολογία του περιβάλλοντος αλλά και των καιρικών συνθηκών. Η αναπαράσταση του χώρου καταστάσεων πρέπει να περιλαμβάνει όλη εκείνη την απαραίτητη πληροφορία που χρειάζεται ένας πράκτορας ώστε να φτάσει στην ανακάλυψη της βέλτιστης πολιτικής. Επιτύχαμε με ελάχιστη πληροφορία στην αναπαράσταση κάθε κατάστασης - τις συντεταγμένες της πλατφόρμας - ένα αξιόλογο αποτέλεσμα. Η μέθοδος προσφέρει επίσης τη δυνατότητα

παραγωγής πολιτικών για διάφορες περιβαλλοντικές συνθήκες. Όπως γίνεται αντιληπτό, το περιβάλλον μεταβάλλεται δυναμικά, γεγονός που καθιστά σημαντική τη δυσκολία του προβλήματος.

Η διατριβή μας δύναται να επεκταθεί μελλοντικά σε πληθώρα τομέων. Αρχικά, μας δίνει τη δυνατότητα να ασχοληθούμε με το μείζον θέμα της εξοικονόμησης ενέργειας θέτοντας περισσότερες δυνάμεις ή προσανατολισμούς (ενέργειες) των κινητήρων. Με αυτό τον τρόπο, η ρομποτική κατασκευή θα είναι σε θέση κάθε χρονική στιγμή να επιλέγει με βάση τις επικρατούσες συνθήκες την ενεργειακά λιγότερο δαπανηρή λύση. Επιπρόσθετα εξαιτίας του μεγάλου χώρου καταστάσεων χρησιμοποιήσαμε μια εκτίμησης της συνάρτησης κατάστασης κι πιτύχαμε ένα αξιόλογο αποτέλεσμα με τη χρήση του γραμμικού μοντέλου. Αναμένεται ότι πιο περίπλοκα περιγραφικά μοντέλα θα βελτιώσουν περισσότερο την απόδοση της μεθόδου. Επίσης, αξίζει να επισημανθεί, πως η χρήση διαφορετικών συναρτήσεων βάσης από τις υπάρχουσες για τη δημιουργία των χαρακτηριστικών καθώς και ο περιορισμός στις τιμές των παραμέτρων των βαρών $w, u$ επιδέχονται μελλοντικές βελτιώσεις. Τέλος, μία σημαντική περιοχή για μελλοντική έρευνα είναι η χρήση περισσότερων χαρτών, είτε τεχνητούς είτε πραγματικούς, κατά την πειραματική αξιόληση της μεθόδου.

# Βιβλιογραφια


[1] M. Carreras, J. Yuh, J. Battle, and P. Ribao, "A behavior-based scheme using reinforcement learning for autonomous underwater vehicles," *IEEE Journal of Ocean Engineering*, vol. 30, pp. 416–427, 2005.

[2] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, (San Francisco, CA, USA), pp. 663–670, Morgan Kaufmann Publishers Inc., 2000.

[3] H. Kawano, "Method for applying reinforcement learning to motion planning and control of under-actuated underwater vehicle in unknown non-uniform sea flow," in *IEEE International conference on Intelligent Robots and Systems (IROS)*, pp. 996–1002, 2005.

[4] G. Antonelli, *Underwater Robots*, vol. 96 of *Springer Tracts in Advanced Robotics*. Springer, 2014.

[5] M. Seto, *Marine Robot Autonomy*. Spinger-Verlag, 2013.

[6] M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning: State-of-the-Art*. Spinger-Verlag, 2012.

[7] S. Hoerner, *Fluid-Dynamic Drag*. Hoerner Publications, 1965.

[8] T. Fossen, *Guidance and Control of Ocean Vehicles*. John Wiley and Sons, 1994.

[9] K. Vlachos and E. Papadopoulos, "Modeling and control of a novel over-actuated marine floating platform," *Ocean Engineering*, vol. 98, pp. 10–22, 2015.

[10] B. Yoo and J. Kim, "Path optimization for marine vehicles in ocean ccurrents using reinforcement learning," *Journal of Marine Science and Technology*, pp. 1–10, 2015.

[11] A. Tsopelakos, K. Vlachos, and E. Papadopoulos, "Backstepping control with energy reduction for an over-actuated marine platform," in *IEEE Intern. Conference on Robotics and Automation (ICRA)*, pp. 553–558, 2015.

[12] N. Tziortziotis, C. Dimitrakakis, and K. Blekas, "Cover tree bayesian reinforcement learning," *Journal of Machine Learning Research*, vol. 15, pp. 2313–2335, 2014.

[13] N. Tziortziotis and K. Blekas, "Model-based reinforcement learning using online clustering," in *IEEE Inern. Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 712–718, 2012.

[14] N. Tziortziotis, C. Dimitrakakis, and K. Blekas, "Linear bayesian reinforcement learning," in *Inern. Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1721–1728, 2013.

[15] L. Busoniu, D. Ernst, B. D. Schutter, and R. Babuska, "Online least-squares policy iteration for reinforcement learning control," in *American Control Conference (ACC)*, pp. 486–491, 2010.

[16] L. Li, M. Littman, and C. Mansley, "Online exploration in least-squares policy iteration," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 733–739, 2009.

[17] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

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

[19] L. Kaelbling, M. Littman, and A. Moore, "Reinforcement learning: A survey," *Journal of Artificial Inteligence Research*, vol. 4, pp. 237–285, 1996.

[20] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.

[21] J. N. Tsitsiklis and R. Sutton, "Asynchronous stochastic approximation and q-learning," pp. 185–202, 1994.

[22] S. Bradtke and A. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, vol. 22, pp. 33–57, 1996.

[23] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *Journal of Machine Learning Research*, vol. 4, pp. 1107–1149, 2003.

[24] G. Konidaris, S. Osentoski, and P. Thomas, "Value function approximation in reinforcement learning using the fourier basis," in *AAAI Conf. on Artificial Intelligence*, pp. 380–385, 2011.

[25] L. Buşoniu, R. Babuška, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010.

# Δημοσιευσεις Συγγραφεα

N. Tziortziotis, K. Tziortziotis, and K. Blekas, " Play Ms. Pac-Man using an advanced reinforcement learning agent " in proceedings of $8^{th}$ Hellenic Conference on Artificial Intelligence (SETN 2014)

K. Tziortziotis, N. Tziortziotis, K.Vlaxos and K. Blekas, " Autonomous navigation of an over-actuated marine platform using reinforcement learning " in proceedings of $9^{th}$ Hellenic Conference on Artificial Intelligence (SETN 2016)

K. Tziortziotis, K.Vlaxos and K. Blekas, " Reinforcement Learning-based Motion Planning of a TriangularFloating Platform under Environmental Disturbances " in proceedings of $24^{th}$ Mediterranean Conference on Control and Automation (MED 2016)

# Σύντομο Βιογραφικο

Ο Κωνσταντίνος Τζιορτζιώτης γεννήθηκε στα Τρίκαλα το 1991. Το 2009 αποφοίτησε απο το Γενικό Λύκειο Πύλης Τρικάλων και εισήχθη στο Τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων από το οποίο αποφοίτησε το 2014. Το ίδιο έτος συνέχισε τις σπουδές του στο Πρόγραμμα Μεπταπτυχιακών Σπουδών του τμήματος Μηχανικών Η/Υ & Πληροφορικής τους Πανεπιστημίου Ιωαννίνων από το οποίο αποφοίτησε τον Ιούλιο του 2016 αποκτώντας ειδίκευση στις "Τεχνολογίες - Εφαρμογές" . Τα ερευνητικά του ενδιαφέροντα εστιάζονται κυρίως στους τομείς της Ρομποτικής, της Μηχανικής Μάθησης, της Ενισχυτικής Μάθησης και της Εξόρυξης Δεδομένων.