

Ατομική Διπλωματική Εργασία

**ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΟΥΡΣΑΣ ΕΞΟΠΛΙΣΜΟΥ
ΟΠΛΩΝ ΜΕ ΤΟ «ΔΙΛΗΜΜΑ ΤΟΥ ΦΥΛΑΚΙΣΜΕΝΟΥ»
ΚΑΙ ΕΚΠΑΙΔΕΥΣΗ ΤΩΝ ΠΡΑΚΤΟΡΩΝ ΜΕ
ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ**

Ευσέβεια Κυριάκου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2014

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

**Μοντελοποίηση Κούρσας Εξοπλισμού όπλων με το «Δίλημμα
του Φυλακισμένου» και Εκπαίδευση των Πρακτόρων με
Ενισχυτική Μάθηση**

Ευσέβεια Κυριάκου

Επιβλέπων Καθηγητής
Δρ. Χρίστος Χριστοδούλου

Η Ατομική αυτή Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση
των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του
Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2014

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον επόπτη μου Δρ. Χρίστο Χριστοδούλου για την καθοδήγηση που μου παρείχε, τον συνεχή ενθουσιασμό και τη σιγουριά που έδειξε για αυτή την εργασία, καθώς και τον διδακτορικό φοιτητή Βασίλη Βασιλειάδη για την συνεχή υποστήριξη και βοήθεια του.

Επίσης, θα ήθελα να ευχαριστήσω θερμά την οικογένεια μου, η οποία πάντα είναι στο πλευρό μου ενθαρρύνοντας με να προχωρώ μπροστά.

Περίληψη

Σε αυτή την διπλωματική εργασία θέλουμε να μελετήσουμε κατά πόσο μπορεί να υπάρξει συμβιβασμός μεταξύ Ελλάδας και Τουρκίας σε μια «Κούρσα Εξοπλισμού όπλων». Οι πράκτορες εκπαιδεύονται με Ενισχυτική Μάθηση χρησιμοποιώντας τον αλγόριθμο Q-learning. Κάνουμε αντιστοίχιση των αμοιβών σύμφωνα με τον πίνακα αμοιβών του «Διλήμματος του Φυλακισμένου» και οι πράκτορες εκπαιδεύονται σε περιβάλλον «Επαναλαμβανόμενου Διλήμματος Φυλακισμένου», ένα παίγνιο γενικού-αθροίσματος. Η αντιστοίχιση των αμοιβών βασίζεται σε ιστορικά γεγονότα από ένα άρθρο που αναφέρουμε αναλυτικά σε επόμενα κεφάλαια. Επίσης για να κάνουμε την αντιστοίχιση αυτή, λαμβάνουμε υπόψη τα έξοδα που κάνουν οι δύο χώρες σε στρατιωτικό εξοπλισμό. Έχουμε τέσσερις φάσεις στις οποίες οι δαπάνες που γίνονται είναι διαφορετικές. Αρχικά έχουμε και τις δύο χώρες να κάνουν μεγάλες σπατάλες σε στρατιωτικό εξοπλισμό, στην συνέχεια έχουμε τις δύο χώρες να ξοδεύουν πολύ λίγα, σε μεταγενέστερη φάση έχουμε την Ελλάδα να ξοδεύει πολύ περισσότερο και σε τελευταία φάση έχουμε την Τουρκία να ξοδεύει λίγο περισσότερο από την Ελλάδα. Το αποτέλεσμα που πήραμε και στις τέσσερις φάσεις ήταν ο συμβιβασμός όμως κάτω από κάποιες προϋποθέσεις. Όταν και οι δύο χώρες βρίσκονταν σε μεγάλες δαπάνες ο συμβιβασμός μεταξύ τους ήταν πολύ δύσκολος, το ίδιο συνέβαινε και όταν η Ελλάδα ξεπερνούσε κατά πολύ την Τουρκία σε στρατιωτικές δαπάνες. Επίσης είδαμε μέσα από πειραματισμό ότι ο μεγάλος ρυθμός έκπτωσης μας έδινε καλύτερα αποτελέσματα και αυτό ήταν λογικό αφού σε τέτοιου είδους παίγνια μας ενδιαφέρει το κοινό καλό και όχι το ατομικό συμφέρον. Επομένως έχοντας υψηλό ρυθμό έκπτωσης πετυχαίνουμε την βέλτιστη ανταμοιβή μακροπρόθεσμα, που είναι ο συμβιβασμός. Όταν όμως και οι δύο χώρες σπαταλούσαν πολύ λίγα σε στρατιωτική δύναμη, είτε η μεταξύ τους διαφορά σε δαπάνες ήταν μικρή, τότε ο συμβιβασμός ήταν πιο εύκολος. Επίσης να πούμε ότι οι δύο παίχτες επιλέγουν την κάθε ενέργεια τους βασιζόμενοι σε μια πιθανότητα ϵ μιας μη άπληστης πολιτικής της ϵ -greedy που δίνει στο παίγνιο περισσότερη εξερεύνηση ώστε οι δύο παίχτες να φτάσουν στην βέλτιστη μακροπρόθεσμη ανταμοιβή.

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή.....	1
Κεφάλαιο 2: Γνωσιολογικό Υπόβαθρο.....	4
2.1 Ενισχυτική Μάθηση.....	4
2.2 Μέθοδοι Ενισχυτικής Μάθησης.....	7
2.3 Θεωρία Παιγνίων.....	13
2.4 Δίλημμα του Φυλακισμένου.....	15
2.5 Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου	19
2.6 Προηγούμενη Εργασία.....	20
Κεφάλαιο 3: Περιγραφή Συστήματος.....	24
3.1 Εισαγωγή.....	24
3.2 Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων.....	26
3.3 Συστατικά Συστήματος.....	31
Κεφάλαιο 4: Αντιστοίχιση «κούρσας Εξοπλισμού» στο «Επαναλαμβανόμενο Δίλημμα Φυλακισμένου και Αποτελέσματα.....	35
4.1 Πρώτη φάση.....	37
4.2 Δεύτερη φάση.....	44
4.3 Τρίτη φάση.....	52
4.4 Τέταρτη φάση.....	60
4.5 Συνένωση Προηγούμενων φάσεων.....	68
Κεφάλαιο 5: Συμπεράσματα και Μελλοντική Εργασία.....	70
5.1 Σύνοψη.....	70
5.2 Συμπεράσματα.....	73
5.3 Σύγκριση με άλλες μελέτες.....	75
5.4 Μελλοντική Εργασία.....	76
Βιβλιογραφία.....	77
Παράρτημα Α: Πηγαίος Κώδικας.....	A-1-

Κεφάλαιο 1

Εισαγωγή

Η κούρσα εξοπλισμού όπλων είναι ένα δίλλημα που απασχολεί τις πλούστες χώρες του κόσμου, αφού τα οπλικά συστήματα κοστίζουν πολύ με παράπλευρα προβλήματα στην αναπτυξιακή στρατηγική μιας χώρας . Από την άλλη πλευρά, οι εξοπλισμοί προσδίδουν αίσθημα ασφάλειας απέναντι σε ανταγωνιστικούς ή εχθρικούς συνασπισμούς. Μέχρι σήμερα δεν έχει καθησυχάσει αυτή η κούρσα εξοπλισμού με συνέπεια πολλές χώρες να δυσχεραίνουν την οικονομία τους σπαταλώντας μεγάλο όγκο χρημάτων σε στρατιωτικό εξοπλισμό.

Ο όρος κούρσα εξοπλισμού όπλων είναι ο ανταγωνισμός μεταξύ των εθνών για υπεροχή στην ανάπτυξη και τη συσσώρευση των όπλων. Αυτή η κούρσα εξοπλισμού είναι ο σημαντικότερος παράγοντας που κρατά ζωντανές τις εντάσεις μεταξύ 2 αντικρουόμενων κρατών στη δική μας περίπτωση της Ελλάδας και της Τουρκίας. Κάθε κράτος με απλά λόγια, πιστεύει ότι με το να αποκτήσει όσο το δυνατό περισσότερο εξοπλισμό σε όπλα, θα πάρει και τη ψηλότερη θέση στη λίστα με τις πιο ισχυρές χώρες.

Η Ενισχυτική Μάθηση (Reinforcement Learning) (EM) ένας κλάδος της Τεχνητής Νοημοσύνης υπήρξε μια ενεργή ερευνητική περιοχή για πολλά χρόνια (Sutton και Barto, 1998)[1]. Η Ενισχυτική Μάθηση σε ένα σύστημα, βοηθάει τους παίκτες μέσα από την αλληλεπίδραση τους με το περιβάλλον να μάθουν έτσι ώστε να επιλέγουν συμπεριφορές που να τους επιφέρουν επιβραβεύσεις και όχι τιμωρίες. Γεγονός που τους δίνει την δυνατότητα μέσα από την εμπειρία να φτάσουν στις βέλτιστες μακροπρόθεσμες ανταμοιβές. Κάθε ενέργεια που εκτελείται από τον παίκτη αξιολογείται θετικά με επιβράβευση είτε αρνητικά με τιμωρία. Στόχος είναι να μάθει από τη αναμενόμενη ανταμοιβή να επιλέγει τις ενέργειες που θα επιφέρουν τη μεγαλύτερη συνολική ανταμοιβή

μακροπρόθεσμα. Υπάρχουν δύο στρατηγικές που μπορεί να ακολουθήσει ένας παίχτης. Η πρώτη είναι η άπληστη πολιτική όπου συνεχώς κάνει εκμετάλλευση της γνώσης του και η άλλη είναι η μη-άπληστη πολιτική στην οποία εξερευνεί. Η εκμετάλλευση θα του επιφέρει μια μεγάλη άμεση ανταμοιβή, όμως η εξερεύνηση ίσως του επιφέρει μια μεγαλύτερη μακροπρόθεσμη συνολική ανταμοιβή. Άρα καλό είναι ο παίχτης να χρησιμοποιήσει την προϋπάρχουσα γνώση του, όμως να κάνει και εξερεύνηση για ενέργειες που θα του επιφέρουν καλύτερες ανταμοιβές στο μέλλον.

Σε προηγούμενα χρόνια η περισσότεροι έκαναν μελέτες EM σε περιβάλλον ενός πράκτορα (single-agent), όμως τον τελευταίο καιρό υπήρξε μεγάλο ενδιαφέρον για το περιβάλλον πολλαπλών πρακτόρων (Hu και Wellman 1998[17], Littman 2001 [12].). (Με τον όρο «πράκτορας» εννοούμε το σύστημα που μαθαίνει). Έτσι έχουμε την Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων (EMΠΠ) (Multiagent Reinforcement Learning problem). Πολλοί πράκτορες εκπαιδεύονται ταυτόχρονα με EM προσπαθώντας να πάρουν την μεγαλύτερη συνολική ανταμοιβή.

Η θεωρία παιγνίων είναι μια μεθοδολογία ανάλυσης καταστάσεων μεταξύ πολλών συμμετεχόντων που έχει σκοπό ο κάθε ένας να αποκτήσει το μεγαλύτερο όφελος μέσα από τον ανταγωνισμό. Μας βοηθάει να καταλάβουμε διάφορες καταστάσεις στις οποίες αλληλεπιδρούν δύο ή περισσότερες οντότητες, κάθε μία από τις οποίες συμπεριφέρεται με στρατηγικό τρόπο και προσπαθεί να πάρει κάποιες αποφάσεις. Η μεμονωμένη οντότητα στην συγκεκριμένη περίπτωση ονομάζεται παίκτης, και είναι αυτός που παίρνει αποφάσεις. Σκοπός του κάθε παίκτη είναι να μεγιστοποιήσει το κέρδος του, το οποίο μετράται σε μια κλίμακα ωφέλειας. Επομένως, η εκμάθηση σε έναν κόσμο πολλαπλών πρακτόρων μπορεί να αντιμετωπισθεί σαν εκμάθηση σε ένα παίγνιο με πολλαπλούς παίκτες.

Σύμφωνα με τα χαρακτηριστικά των ανταμοιβών του κάθε παίχτη έχουμε 2 κατηγορίες παιγνίων. Όταν η επιβράβευση ενός παίκτη είναι ίση με την τιμωρία του αντιπάλου του, το παίγνιο ονομάζεται “παίγνιο μηδενικού αθροίσματος” (zero-sum games). Σε αυτά τα

παίγνια, το άθροισμα των αμοιβών είναι ίσο με μηδέν με αποτέλεσμα η συνεργασία για τους παίκτες να είναι ανέφικτη. Αντίστοιχα υπάρχουν “παίγνια μη-μηδενικού αθροίσματος” (non zero-sum game) στα οποία το άθροισμα των αμοιβών είναι διάφορο του μηδενός. Το κέρδος κάποιου δεν σημαίνει απαραίτητα τη ζημιά κάποιου ανταγωνιστή, και οι δύο μπορεί να κερδίσουν ή και να χάσουν αντίστοιχα.

Μια κατηγορία της θεωρίας παιγνίων είναι το Δίλημα του Φυλακισμένου (ΔΦ). Το παίγνιο αυτό ανήκει στη κατηγορία παιγνίου γενικού αθροίσματος για τον λόγο ότι δεν θα υπάρξει νικητής. Με την χρήση Ενισχυτικής Μάθησης στο ΔΦ ευελπιστούμε ότι τα δύο κράτη θα μάθουν πως να φτάσουν σε συμβιβασμό. Αυτός είναι και ο απώτερος σκοπός αυτής της εργασίας. Με την εισαγωγή της μάθησης στο πρόβλημα της κούρσας εξοπλισμού, απαιτείται τα δύο κράτη να παίζουν πολλούς γύρους, ώστε να μάθουν να φτάνουν σε αμοιβαίο συμβιβασμό. Γι’ αυτό τον λόγο θα χρησιμοποιήσουμε το «επαναλαμβανόμενο Δίλημα του Φυλακισμένου» και την Ενισχυτική Μάθηση για την εκπαίδευση των δυο κρατών Ελλάδας και Τουρκίας.

Οι επιλογές των δύο κρατών είναι να δουν το κοινό καλό και να συνεργαστούν είτε να κοιτάζουν το προσωπικό τους συμφέρον με αποτέλεσμα με το να εξοπλίζονται ασύστολα. Το προσωπικό συμφέρον θα επιφέρει την μέγιστη ανταμοιβή σε ατομικό επίπεδο σε αντίθεση με τον αντίπαλο του που θα προτιμήσει να συνεργαστεί και θα πάρει την χειρότερη ανταμοιβή, όμως αν και οι δυο κοιτάζουν το προσωπικό τους συμφέρον τότε θα πάρουν και οι δύο σχετικά μικρή ανταμοιβή. Ενώ στη περίπτωση που και οι δύο επιλέξουν να συνεργαστούν θα πάρουν μια μέτρια ανταμοιβή. Επίσης να έχουμε υπόψη μας ότι ο κάθε παίχτης δεν γνωρίζει την επιλογή του αντιπάλου του σε μια δεδομένη χρονική στιγμή.

Θα χρησιμοποιήσω τον αλγόριθμο Q-learning (αλγόριθμος EM) για την εκμάθηση των δύο πρακτόρων ξεχωριστά πάνω στο μοντέλο του ΔΦ και επίσης θα προσθέσω μηχανισμό εξερεύνησης ε-greedy για την περεταίρω εξερεύνηση του κάθε πράκτορα ως προς την επιλογή της ενέργειας του.

Κεφάλαιο 2

Γνωσιολογικό Υπόβαθρο

2.1 Ενισχυτική Μάθηση	4
2.2 Μέθοδοι Ενισχυτικής Μάθησης	7
2.3 Θεωρία Παιγνίων	13
2.4 Δίλημμα του Φυλακισμένου	15
2.5 Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου	19
2.6 Προηγούμενη Εργασία	20

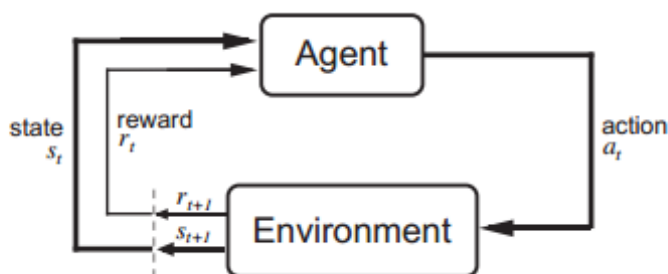
2.1 Ενισχυτική Μάθηση

Ενισχυτική μάθηση είναι μια προσέγγιση της Τεχνητής Νοημοσύνης που δίνει έμφαση στη μάθηση από την ατομική μορφή που αλληλεπιδρά με το περιβάλλον της [3]. Επίσης είναι η εκμάθηση του τι πρέπει να κάνουμε, πώς δηλαδή να χαρτογραφήσουμε τις διάφορες καταστάσεις με τις ενέργειες έτσι ώστε να μεγιστοποιήσουμε την συνολική ανταμοιβή. Στον εκπαιδευόμενο δεν δίνονται κατευθυντήριες γραμμές για το τι ενέργειες να επιλέξει, όπως σε άλλες μορφές μηχανικής μάθησης, αλλά αντίθετα πρέπει να ανακαλύψει από μόνος του ποιές ενέργειες θα του δώσουν τη μεγαλύτερη ανταμοιβή δοκιμάζοντας τις.

Στην Ενισχυτική Μάθηση υπάρχουν δύο στρατηγικές η άπληστη που εκμεταλλεύεται την γνώση που προϋπάρχει σε ένα σύστημα και η μη-άπληστη στρατηγική, στην οποία ο παίχτης κάνει εξερεύνηση. Η εκμετάλλευση θα επιφέρει μεγάλη άμεση ανταμοιβή όμως η εξερεύνηση ίσως δώσει μια μεγαλύτερη συνολική ανταμοιβή μακροπρόθεσμα.

Ο ΠΡΑΚΤΟΡΑΣ είναι ο εκπαιδευόμενος, που πρέπει μέσα σε άγνωστες περιοχές να είναι ικανός να μαθαίνει μέσα από την εμπειρία του, για να πάρει τη μεγαλύτερη ανταμοιβή. Το ΠΕΡΙΒΑΛΛΟΝ μπορούμε να τον χαρακτηρίσουμε ως ένα «κριτή» ο οποίος δίνει τιμωρίες και αμοιβές αξιολογώντας τη κάθε ενέργεια που κάνει ο πράκτορας. ΕΝΕΡΓΕΙΑ είναι μια κίνηση που μπορεί να κάνει ο πράκτορας σε μια δεδομένη κατάσταση. Ένα ΕΠΕΙΣΟΔΙΟ μπορεί να οριστεί ως μια περίπτωση επιλογής μιας ενέργειας. Η ΚΑΤΑΣΤΑΣΗ φυλάγεται στο σύστημα για να αναπαραστήσει μια προηγούμενη συμπεριφορά και να χρησιμοποιηθεί για να καθορίσει την επόμενη συμπεριφορά. Και τέλος το ΕΝΙΣΧΥΤΙΚΟ ΣΗΜΑ αξιολογεί την ενέργεια που συμβαίνει τώρα, χωρίς να λέει στον εκπαιδευόμενο ποια είναι η βέλτιστη ενέργεια για να επιλέξει. Η ΣΥΝΑΡΤΗΣΗ ΑΞΙΑΣ είναι η διαδικασία που παράγει ένα άμεσο ενισχυτικό σήμα σε κάθε ενέργεια που επιλέγει ο πράκτορας από κάθε κατάσταση. Αυτή η συνάρτηση δεν βρίσκεται μέσα στα όρια ελέγχου του πράκτορα, οπότε υπάρχει έξω απ' αυτόν (στο περιβάλλον). Ο στόχος είναι η μεγιστοποίηση της αναμενόμενης ολικής αμοιβής μετά από κάποια χρονική περίοδο.

ΜΟΝΤΕΛΟ ΕΝΥΣΧΥΤΙΚΗΣ ΜΑΘΗΣΗΣ (μέσα από μακροβιανή διαδικασία απόφασης ΜΔΑ)



Εικόνα 2.1

Στο πιο πάνω διάγραμμα έχουμε μακροβιανή διαδικασία απόφασης, που υποστηρίζει μια ιδιότητα που ονομάζεται MARKOV. Με αυτή την ιδιότητα επιτρέπουμε στη κατάσταση να κρατά πληροφορίες που έχουν να κάνουν με προηγούμενες καταστάσεις. Έχοντας μια τέτοια ιδιότητα ο πράκτορας τους συστήματος θα μπορεί να προβλέψει την επόμενη κίνηση του καθώς και την αμοιβή που θα πάρει από την τρέχουσα κατάσταση.

Ο πράκτορας και το περιβάλλον αλληλεπιδρούν σε μια σειρά από διακριτές χρονικές στιγμές $t=1,2,3..$ Σε κάθε βήμα ο πράκτορας διορίζει το περιβάλλον να βρίσκεται σε μια κατάσταση, s_t , και επιλέγει μια ενέργεια a_t . Ως ανταπόκριση το περιβάλλον κάνει τη στοχαστική μεταφορά σε μια νέα κατάσταση, s_{t+1} , και στοχαστικά δίνει μια αριθμητική ανταμοιβή, $r_{t+1} \in \mathcal{R}$. (Εικόνα 2.1). Ο πράκτορας στοχεύει να μεγιστοποιήσει την ανταμοιβή που λαμβάνει μακροπρόθεσμα. Για παράδειγμα ο πιο κοινό σκοπός είναι να επιλέγει κάθε ενέργεια a_t , έτσι ώστε να μεγιστοποιεί την αναμενόμενη επιστροφή έκπτωσης:

$$E\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots\}$$

όπου το γ είναι ο ρυθμός έκπτωσης ($0 \leq \gamma \leq 1$) και r είναι η αμοιβή σε κάθε χρονική στιγμή. Το αναμενόμενο αποτέλεσμα των καθυστερημένων ανταμοιβών χρησιμοποιείται στην Συνάρτηση Αξίας για να καταστήσει ικανό τον πράκτορα να υπολογίζει ή να προβλέπει ποια ενέργεια να κάνει. Η συνάρτηση αξίας έχει δύο τύπους:

- i. *1. $V\pi(s)$: Μπορούμε να καθορίσουμε τη αξία του να βρίσκεσαι σε μια κατάσταση s κάτω από την πολιτική Π*
- ii. *2. $Q\pi(s,a)$: Εφαρμόζει απευθείας την εμπειρία του πράκτορα αλληλεπιδρώντας με το περιβάλλον αλλάζοντας την πολιτική σε πραγματικό χρόνο. Είναι η αξία της επιλογής μιας ενέργειας a στην κατάσταση s ενόσω ακολουθεί μια δοθείσα πολιτική π .*

Κάθε αλγόριθμος έχει την δική του συνάρτηση αξίας. Θα παρουσιάσουμε κάποιες μεθόδους EM παρακάτω, μαζί με τις εξισώσεις για τους πιο κύριους αλγόριθμους που χρησιμοποιούνται.

2.2 Μέθοδοι Ενισχυτικής Μάθησης

Υπάρχουν τρεις κύριες κλάσεις μεθόδων EM: Δυναμικός Προγραμματισμός (ΔΠ), Monte Carlo και μάθηση Χρονικών Διαφορών (Temporal Difference). Υπάρχουν πλεονεκτήματα αλλά και μειονεκτήματα και στις τρεις παρόλα αυτά ακολουθούν τις ίδιες διαδικασίες: αξιολόγηση πολιτικής (policy evaluation) που είναι εκτίμηση της Συνάρτησης Αξίας και η βελτίωση πολιτικής (policy improvement), που ενημερώνει τη Συνάρτηση Αξίας και βελτιώνει την πολιτική του πράκτορα.

Ο ΔΠ έχοντας ως δεδομένο ένα τέλειο μοντέλο περιβάλλοντος με την μορφή MDA μπορεί να υπολογίζει τις βέλτιστες πολιτικές. Πατέρας του δυναμικού προγραμματισμού θεωρείται ο Richard Bellman 1957 [19]. Ο Bellman με τον όρο δυναμικός προγραμματισμός ονόμασε τον τρόπο με τον οποίο επιλύονται τα προβλήματα στα οποία πρέπει να ληφθούν μία σειρά αποφάσεων που η κάθε μια από αυτές επηρεάζει τις επόμενες αποφάσεις, και που όλες μαζί δημιουργούν ένα βέλτιστο αποτέλεσμα.

Η μέθοδος του ΔΠ για να βρει την βέλτιστη πολιτική χρησιμοποιεί Συναρτήσεις Αξίας. Χρησιμοποιεί εξισώσεις Bellman (1957) για την προσέγγιση επιθυμητών Συναρτήσεων Αξίας. Αυτό μπορεί να το κάνει μέσω 2 μεθόδων: αξιολόγηση πολιτικής και βελτίωση πολιτικής.

Αξιολόγηση της πολιτικής διαφορετικά γνωστή ως το πρόβλημα της πρόβλεψης (prediction problem) είναι ο υπολογισμός της συνάρτησης κατάστασης-αξίας ($V(s)$). Αυτή η συνάρτηση αξίας υπολογίζεται λαμβάνοντας υπόψη τις εκτιμήσεις για τις επόμενες καταστάσεις, η οποίες πρέπει να φυλάγονται. Αυτό έχει αρνητικά αποτελέσματα αφού είναι υπολογιστικά ακριβό. Συνήθως για να τερματιστεί η αξιολόγηση πολιτικής πρέπει να υπάρχει μικρή διαφορά μεταξύ της τρέχουσας και της προηγούμενης συνάρτησης αξίας.

Βελτίωση Πολιτικής είναι το πως να βρεθεί μια καλύτερη πολιτική από την αρχική που θα επιφέρει μεγαλύτερο αναμενόμενο κέρδος. Η επανάληψη πολιτικής (policy iteration) συνδυάζει αξιολόγηση και βελτίωση με την περικοπή των δύο ξεχωριστών διαδικασιών για να γίνει πιο φθηνή υπολογιστικά.

Το πιο σοβαρό μειονέκτημα του ΔΠ είναι ότι απαιτεί ένα πλήρες και ακριβές μοντέλο του περιβάλλοντος (οποίο δεν είναι πάντα διαθέσιμο). Επίσης, η μέθοδος του ΔΠ χρειάζεται όλες τις καταστάσεις και ενέργειες έτσι είναι υπολογιστικά ακριβής σε σύγκριση με άλλες μεθόδους ΕΜ.

Η μέθοδος Monte Carlo δεν χρειάζεται να γνωρίζει πλήρως το περιβάλλον [4]. Μαθαίνει μέσα από την εμπειρία και είναι παρόμοια με την μέθοδο του Δυναμικού Προγραμματισμού. Επίσης και αυτή η μέθοδος βασίζεται στην εκτίμηση της Συνάρτησης Αξίας. Για να δώσει αξία σε μια πολιτική υπολογίζει τον μέσο όρο των αποτελεσμάτων σε κάποιες καταστάσεις, και σταδιακά ο μέσος όρος για κάποια κατάσταση θα πρέπει να συγκλίνει, δηλαδή να μην αλλάξει. Δεν κάνει bootstrapping έτσι έχουμε λιγότερη ζημιά αν υπάρξει παράβαση. Αυτή η πολιτική είναι λιγότερο απαιτητική από το Δυναμικό προγραμματισμό γιατί όπως είπαμε δεν χρειάζεται ένα πλήρες μοντέλο περιβάλλοντος για να μάθει αλλά μαθαίνει κατευθείαν από το περιβάλλον μέσω δειγματοληψίας, συγκρίνει κάποιες καταστάσεις με ένα σύνολο καταστάσεων για να διαλέξει την πιο βέλτιστη.

Για να μπορεί να συνεχίσει την εξερεύνηση νέων ενεργειών, σε σύγκριση με την εκμετάλλευση γνωστών ενεργειών οι οποίες αποφέρουν αμοιβές, η μέθοδος Monte Carlo χρησιμοποιεί είτε ένα αλγόριθμο εντός-πολιτικής (on-policy), ο οποίος προσπαθεί να αξιολογήσει ή να βελτιώσει την πολιτική που χρησιμοποιείται για τη λήψη καλύτερων αποφάσεων, ή ένα αλγόριθμο εκτός-πολιτικής (off-policy), ο οποίος περιλαμβάνει μια πολιτική εκτίμησης που βελτιώνει μια ενέργεια και μια πολιτική συμπεριφοράς κάνει τη βελτίωση.

Η μέθοδος *Χρονικών Διαφορών (ΧΔ)* συμπεριφέρεται όπως η μέθοδος Monte Carlo επειδή δεν απαιτεί ένα πλήρες μοντέλο του περιβάλλοντος, αλλά μόνο εμπειρία [5]. Επίσης έχει την βάση του σε προηγούμενες μαθημένες εμπειρίες για να μπορεί να εκτιμήσει όπως ακριβώς και στο Δυναμικό Προγραμματισμό. Χρησιμοποιούμε εμπειρία για ενημέρωση των εκτιμήσεων της Συνάρτησης Αξίας όπως φαίνεται πιο κάτω.

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

Εικόνα 2.2

Όπου $V(S_t)$ είναι η συνάρτηση κατάστασης-αξίας της κατάστασης S σε χρόνο t , α είναι ο ρυθμός μάθησης, r_{t+1} η αμοιβή που δόθηκε τη χρονική στιγμή $t+1$, γ είναι ο ρυθμός έκπτωσης και $V(S_{t+1})$ είναι η συνάρτηση κατάστασης-αξίας της κατάστασης S σε χρόνο $t+1$. Η αξία της προηγούμενης κατάστασης ενημερώνεται με βάση τη νέα κατάσταση, δηλαδή γίνεται bootstrapping και χρειάζεται να περιμένει μόνο μέχρι το τέλος μιας χρονικής στιγμής για να ενημερώσει τη Συνάρτηση Αξίας, ενώ η μέθοδος Monte Carlo χρειάζεται να περιμένει μέχρι το τέλος ενός επεισοδίου.

Υπάρχουν διάφορες μέθοδοι Χρονικών Διαφορών. Ο αλγόριθμος Q-Learning (Watkins, 1989 [18]) δεν δίνει προσοχή στην πολιτική που ακολουθείται [7]. Αντίθετα, χρησιμοποιεί μόνο τη καλύτερη Q-Value τιμή. Έτσι, είναι ένας αλγόριθμος εκμάθησης εκτός-Πολιτικής. Λέγεται εκτός-πολιτικής, διότι η πολιτική που έμαθε μπορεί να είναι διαφορετική από ό, τι η πολιτική που εκτελέστηκε. Συγκεκριμένα, μαθαίνει την άπληστη πολιτική ενόσω χρησιμοποιεί μια πολιτική που εξερευνά, δηλαδή επιλέγει μη-άπληστες ενέργειες. Ακολουθεί η εξίσωση του αλγορίθμου Q-Learning:

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Εικόνα 2.3

Όπου $Q(s,a)$ είναι η τιμή της συνάρτησης αξίας, και οι θέσεις του πίνακα απαρτίζονται από το s που είναι η κατάσταση στην οποία βρίσκεται ο πράκτορας και επιλέγει την ενέργεια a . Επίσης όπου α είναι ο ρυθμός μάθησης και r η άμεση ανταμοιβή που δίνεται στο πράκτορα. Το γ είναι ο ρυθμός έκπτωσης και το $\max_a Q(s',a')$ είναι η μέγιστη τιμή Q που υπολογίζετε από όλες τις πιθανές ενέργειες που αντιστοιχούν στην επόμενη κατάσταση.

Η μέθοδος «ενέργειας-κριτικής» (actor-critic method) είναι μέθοδος ΧΔ η οποία έχει ξεχωριστή δομή μνήμης για να παρουσιάζει ξεκάθαρα την πολιτική ανεξάρτητα από τα την συνάρτηση αξίας [6]. Η δομή της πολιτικής είναι γνωστή ως ηθοποιός (actor), επειδή χρησιμοποιείται για να επιλέξει ενέργειες και η υπολογιζόμενη συνάρτηση αξίας είναι γνωστή ως κριτής (critic), διότι κρίνει της ενέργειες που γίνονται από τον ηθοποιό. Είναι μια εντός-πολιτικής μέθοδος Χρονικής Διαφοράς, αφού ο κριτής πρέπει να μάθει το ενισχυτικό σήμα του ηθοποιού. Πριν την ανάπτυξη του αλγορίθμου Q-learning, η μέθοδος ενέργειας-κριτικής ήταν η μέθοδος που προτιμούσαν αρκετοί ερευνητές του τομέα της ενισχυτικής μάθησης.

Η μέθοδος ΧΔ είναι η πιο διαδεδομένη μέθοδος ενισχυτικής μάθησης, αφού είναι απλή, λειτουργεί βήμα με βήμα και είναι υπολογιστικά λιγότερο ακριβή από άλλες μεθόδους ΕΜ. Χρησιμοποιώντας τη μέθοδο ΧΔ μπορείς να μάθεις πριν γνωρίσεις την τελική ανταμοιβή πώς να κατανέμεις ανταμοιβές (credit) στις διάφορες ενέργειες. Αυτό είναι γνωστό ως το πρόβλημα της Χρονικής Απόδοσης ανταμοιβών (Temporal Credit Assignment problem). Στις μεθόδους Χρονικών Διαφορών, το σφάλμα δ_t υπολογίζεται ανάλογα με την παρακάτω εξίσωση και προσαρμόζεται αναλόγως στις κατάλληλες καταστάσεις:

$$\delta_t = r_{t+1} + \gamma \cdot V(S_{t+1}) - V(S_t)$$

Εικόνα 2.4

Το t συμβολίζει την τρέχουσα χρονική στιγμή, r_{t+1} είναι η ανταμοιβή (ενισχυτικό σήμα) τη δεδομένη χρονική στιγμή $t+1$, καθώς το γ είναι ο ρυθμός έκπτωσης των ανταμοιβών που θα έρθουν στο μέλλον, $V(S_{t+1})$ είναι η συνάρτηση κατάστασης-αξίας της χρονικής στιγμής $t+1$ και $V(S_t)$ η συνάρτηση κατάστασης-αξίας της χρονικής στιγμής t .

Το ίχνος εκλεξιμότητας (eligibility trace) λ είναι ένας από τους βασικούς μηχανισμούς της ενισχυτικής μάθησης [9]. Σχεδόν κάθε μέθοδος ΧΔ, όπως το Q-learning ή SARSA, μπορεί να συνδυαστεί με ίχνος εκλεξιμότητας για να αποκτήσουν μια πιο γενική μέθοδο που να μπορεί να μάθει πιο αποτελεσματικά. Επίσης το ίχνος εκλεξιμότητας βοηθάει στο να επιταχυνθεί η διαδικασία της διάδοσης ανταμοιβής στην EM και είναι ο βασικός μηχανισμός για χρονική απόδοση ανταμοιβής.

Υπάρχουν δύο τρόποι για να χρησιμοποιείται το ίχνος εκλεξιμότητας. Η πιο θεωρητική όψη, είναι η ύπαρξη μιας γέφυρας από ΧΔ στις μεθόδους Monte Carlo. Όταν οι μέθοδοι ΧΔ είναι επαυξημένες με ίχνος εκλεξιμότητας, παράγουν μία οικογένεια μεθόδων που εκτείνονται σε ένα φάσμα που έχουν τις μεθόδους Monte Carlo στο ένα άκρο και μεθόδους ΧΔ στο άλλο άκρο. Στο ενδιάμεσο υπάρχουν άμεσες μέθοδοι που είναι καλύτερες συνήθως από τις ακραίες μεθόδους. Υπό αυτή την έννοια το ίχνος εκλεξιμότητας σε αυτές τις μεθόδους κοστίζει και είναι αποκαλυπτικό.

Ο άλλος τρόπος είναι το ίχνος εκλεξιμότητας να είναι μια προσωρινή καταγραφή της εμφάνισης ενός γεγονότος, όπως η επίσκεψη σε μια κατάσταση ή η επιλογή μιας ενέργειας [8]. Η βασική ιδέα είναι ότι όταν συμβαίνει ένα σφάλμα, μόνο στις τρέχουσες καταστάσεις αποδίδεται ευθύνη γι' αυτό ενώ σε καταστάσεις που έγιναν προηγουμένως δίνεται λιγότερη ευθύνη. Ένα ίχνος εκλεξιμότητας μπορεί να αυξάνεται κάθε φορά που γίνεται επίσκεψη μιας κατάστασης και μετά να εξαφανίζεται βαθμιαία όταν η κατάσταση δεν επισκέπτεται. Η ακόλουθη εξίσωση υπολογίζει το ίχνος εκλεξιμότητας για μια δεδομένη κατάσταση, $e_t(s)$. Αν η κατάσταση έχει επισκεφθεί, τότε το ίχνος εκλεξιμότητάς της αυξάνεται κατά 1, αλλιώς εξασθενεί με την πάροδο του χρόνου:

$$e_t(s) = \begin{cases} \gamma \lambda e_{t-1}(s), & s \neq s_t \\ \gamma \lambda e_{t-1}(s) + 1, & s = s_t \end{cases}$$

Εικόνα 2.5

Το γ συμβολίζει τον ρυθμό έκπτωσης, το λ είναι η παράμετρος εξαφάνισης ίχνους με τιμές που κυμαίνονται μεταξύ 0 (που αντιπροσωπεύει bootstrapping) και 1 (που αντιπροσωπεύει nonbootstrapping), $e_t(s)$ είναι το ίχνος εκλεξιμότητας της κατάστασης s τη χρονική στιγμή t . Η Εξίσωση 2.6 υπολογίζει την αλλαγή στην συνάρτηση κατάστασης-αξίας $V(s)$ για τις πρόσφατα επισκεφθείσες καταστάσεις:

$$\Delta V(s) = a \cdot \delta_t \cdot e_t(s)$$

Εικόνα 2.6

όπου a είναι ο ρυθμός μάθησης, δ_t είναι το σφάλμα χρονικών διαφορών που δίνεται από την εξίσωση 2.4 και $e_t(s)$ είναι το ίχνος εκλεξιμότητας της κατάστασης s τη χρονική στιγμή t που δίνεται από την εξίσωση 2.5.

2.3 Θεωρία Παιγνίων

Η Θεωρία Παιγνίων είναι ο επιστημονικός κλάδος που μελετάει συστηματικά με την χρήση μαθηματικών εργαλείων την συμπεριφορά των ατόμων σε συνθήκες στρατηγικής αλληλεπίδρασης. Στρατηγική αλληλεπίδραση υπάρχει όταν το τελικό αποτέλεσμα για κάθε άτομο που παίρνει αποφάσεις δεν εξαρτάται μόνο από τις δικές του επιλογές (ή από την τύχη) αλλά και από τις επιλογές άλλων με τους οποίους το άτομο αυτό συνεργάζεται ή βρίσκεται σε σύγκρουση.

Κύρια στοιχεία ενός παιγνίου είναι: α) οι παίκτες, β) οι κανόνες, γ) οι στρατηγικές των παικτών, δ) η πληροφόρηση που έχει κάθε παίκτης όταν κληθεί να λάβει κάποια απόφαση, και ε) οι αποδόσεις (αμοιβές). Παίκτης ονομάζεται κάθε ένας ο οποίος παίρνει αποφάσεις στη διάρκεια του παιγνίου[2]. Οι κανόνες καθορίζουν θέματα όπως η σειρά με την οποία οι παίκτες λαμβάνουν αποφάσεις και οι συνθήκες κάτω από τις οποίες τερματίζεται ή συνεχίζεται το παίγνιο. Οι στρατηγικές των παικτών είναι ολοκληρωμένα σχέδια δράσεις που εξειδικεύουν τις επιλογές τους σε κάθε σημείο του παιγνίου. Οι παίκτες όταν έλθει η σειρά τους να κάνουν μια επιλογή μπορεί να γνωρίζουν πλήρως την ιστορία του παιγνίου (τι προηγήθηκε μέχρι εκείνη την στιγμή) μπορεί όμως και να μην την γνωρίζουν. Επίσης, μπορεί να γνωρίζουν με ακρίβεια τις αποδόσεις των άλλων παικτών μπορεί όμως και όχι. Οι αποδόσεις προκύπτουν για τους παίκτες με τον τερματισμό του παιγνίου και μπορεί να είναι χρηματικές απολαβές ή επίπεδα ικανοποίησης/χρησιμότητας.

Υποτίθεται ότι όλοι οι παίκτες είναι ορθολογικοί με την έννοια ότι έχουν τη δυνατότητα: α) να κατατάξουν τις αποδόσεις σε περισσότερο και λιγότερο επιθυμητές και β) να προσδιορίσουν τις επιλογές που οδηγούν στην πλέον επιθυμητή απόδοση [10]. Επιπλέον, κάθε παίκτης ξεχωριστά γνωρίζει ότι και οι άλλοι είναι ορθολογικοί και λαμβάνει το στοιχείο αυτό υπόψη του όταν κάνει τις δικές του επιλογές. Το αποτέλεσμα του παιγνίου καθορίζεται από τις επιλογές όλων των παικτών και η λύση του παιγνίου είναι η βέλτιστη στρατηγική όλων των παιχτών.

Η περισσότερη έρευνα στην εφαρμογή της Ενισχυτικής Μάθησης στο πεδίο πολλαπλών πρακτόρων, αφορά καταστάσεις όπου υπάρχει μόνο ένας νικητής ή γενικά οι απολαβές ή οι απώλειες κάποιου παίχτη είναι απόλυτα ισορροπημένες με τις απώλειες ή τις απολαβές των υπόλοιπων παιχτών. Αυτές οι καταστάσεις μοντελοποιούνται ως παίγνια μηδενικού αθροίσματος. Υπάρχουν όμως και οι καταστάσεις, όπου η απολαβή κάποιου παίχτη δεν αντιστοιχεί απαραίτητα στην απώλεια κάποιου άλλου κατά συνέπεια σε αυτά τα παίγνια δεν υπάρχει νικητής ή ηττημένος. Αυτή η μοντελοποίηση αποτελεί παίγνιο γενικού αθροίσματος. Καθώς επίσης αυτό το μοντέλο μπορεί να θεωρηθεί αρκετά πιο δύσκολο αφού απαιτεί ένα μίγμα συνεργασίας και συμβιβασμού.

Η Θεωρία Παιγνίων αναγνωρίζει ότι κάθε παίκτης ξεχωριστά προσπαθεί να προβλέψει τι θα κάνουν οι άλλοι παίκτες και να αντιδράσει με τον καλύτερο γι' αυτόν τρόπο. Στη λύση ενός παιγνίου οι προσδοκίες κάθε παίκτη σχετικά με την συμπεριφορά των υπολοίπων αποδεικνύονται ορθές και οι αντιδράσεις με βάση τις προσδοκίες αυτές είναι άριστες. Η λύση ονομάζεται ισορροπία επειδή σε αυτή (με δεδομένες τις στρατηγικές των υπολοίπων) κανείς από τους παίκτες δεν έχει κίνητρο να αλλάξει τη δική του στρατηγική.

Η θεωρία παιγνίων μπορεί να εφαρμοστεί σε πολλές καταστάσεις όπως εταιρείες που ανταγωνίζονται, πολιτικοί που διεκδικούν ψήφους, ένορκοι που αποφασίζουν για μια ετυμηγορία, ζώα που μάχονται για λεία, παίκτες που συμμετέχουν σε έναν πλειστηριασμό και σε άλλες πολλές καταστάσεις.

2.4 Δίλημμα του Φυλακισμένου

Το «Δίλημμα του Φυλακισμένου» είναι ένα απλό, αλλά χαρακτηριστικό παράδειγμα της θεωρίας παιγνίων και είναι μέρος των παιγνίων γενικού αθροίσματος [11]. Υπάρχουν 3 διαφορετικές καταστάσεις που μπορεί να προκύψουν: Να βγουν και οι 2 κερδισμένοι με ένα αμοιβαίο συμβιβασμό, να βγουν και οι 2 ηττημένοι με την επιμονή να κρατήσουν αυτή την κούρσα μεταξύ τους καθώς και ο ένας να γίνει το κορόιδο του άλλου παίρνοντας την μικρότερη αμοιβή έναντι της καλύτερης που θα πάρει ο αντίπαλος του. Δεν υπάρχει καμία επικοινωνία μεταξύ των παιχτών, κανένας από τους 2 δεν γνωρίζει τι ενέργεια θα επιλέξει ο αντίπαλος του. Επίσης οι ενέργειες των 2 παιχτών επιλέγονται ταυτόχρονα. Ανάλογα με την ενέργεια που θα επιλέξει ο κάθε παίχτης σε συνάρτηση με την ενέργεια του αντιπάλου θα πάρει και την ανάλογη ανταμοιβή.

Στη δική μας περίπτωση στο ΔΦ έχουμε 2 κράτη Ελλάδα, Τουρκία που πρέπει να αποφασίσουν εάν θα συμβιβαστούν (να σταματήσουν δηλαδή τις μεγάλες σπατάλες σε στρατιωτικό εξοπλισμό) ή να συνεχίσουν να ανταγωνίζονται. Σύμφωνα με την ορολογία της θεωρίας των παιγνίων, μπορούμε να πούμε ότι κάθε κράτος έχει δύο επιλογές στρατηγικής: την αγορά στρατιωτικού εξοπλισμού-μεγάλη σπατάλη χρημάτων (b) και την μη αγορά στρατιωτικού εξοπλισμού(n). Δύο κυβερνήσεις με δύο επιλογές στρατηγικής παράγουν τον ακόλουθο πίνακα:

Προϋποθέσεις:

1. $T > R > P > S$
2. $2R > T + S$

REWARD-R, SUCKER'S PAYOFF-S, TEMPTATION-T, PUNISHMENT-P

		Ελλάδα	
		Not build	Build
Τουρκία	Not build	(n,n) R(3,3)	(n,b) S(1) T(4)
	build	(b,n) T(4) S(1)	(b,b) P(2,2)

Η Τουρκία αντιπροσωπεύει την πρώτη τιμή σε κάθε τετράγωνο, ενώ η Ελλάδα αντιπροσωπεύει την δεύτερη τιμή στο κάθε τετράγωνο.

Στο πρώτο κελί πάνω αριστερά έχουμε τον συνδυασμό στρατηγικής (n,n), που σημαίνει αν η Τουρκία επιλέξει να μην αγοράσει εξοπλισμό και η Ελλάδα δεν αγοράσει εξοπλισμό τότε τα 2 κράτη δεν θα εμπλακούν σε κούρσα εξοπλισμών. Οι αμοιβές για τα 2 κράτη που επέλεξαν την ίδια ενέργεια είναι οι ίδιες. Την έθεσα 3 γιατί πρέπει να είναι η μεγαλύτερη συνολική αμοιβή αφού θέλουμε να φτάσουμε σε κάποιο συμβιβασμό, και επίσης να είναι μεγαλύτερη από την τιμή πιο διάλεξα για την αγοράσουν εξοπλισμό του να αγοράσουν που θα δούμε στην 3η περίπτωση.

(Κελί πάνω δεξιά) Αν η Τουρκία επιλέξει να μην αγοράσει εξοπλισμό ενώ η Ελλάδα επιλέξει να αγοράσει εξοπλισμό (n,b), τότε η Ελλάδα αποκτά ένα πλεονέκτημα δύναμης

πάνω στην Τουρκία. Έθεσα τιμή 4 αυτή που παίρνει το πλεονέκτημα δύναμης (είναι αυτή με τη μεγαλύτερη ατομική αμοιβή). Και δήλωσα τιμή 1 το κράτος που γίνεται κορόιδο και παίρνει την χειρότερη αμοιβή. Συνολικά πρέπει να βγαίνει μια μικρότερη αμοιβή από την προηγούμενη περίπτωση.

(κελί κάτω αριστερά) Αν η Τουρκία αποφασίσει να αγοράσει εξοπλισμό και το ίδιο αποφασίσει και η Ελλάδα (b,b), τότε και τα 2 κράτη θα εμπλακούν σε μια κούρσα εξοπλισμών. Οι αμοιβές για τα 2 κράτη που επέλεξαν την ίδια ενέργεια είναι η ίδια. Την έθεσα 2 γιατί συνολικά πρέπει να είναι μικρότερη από την πρώτη περίπτωση συμβιβασμού. Όσο αφορά τη κάθε τιμή ξεχωριστά, πρέπει να είναι μικρότερη από την αμοιβή το κράτους που έχει το πλεονέκτημα της δύναμης (<4) και μεγαλύτερη από την τιμή που έχει το κράτος που έγινε κορόιδο.

Τέλος αν η Τουρκία αποφασίσει να αγοράσει εξοπλισμό, ενώ η Ελλάδα δεν θα αγοράσει εξοπλισμό (b,n), τότε η Τουρκία αποκτά ένα πλεονέκτημα δύναμης πάνω στην Ελλάδα. Για τους λόγο που εξήγησα στη δεύτερη περίπτωση έθεσα 4 στην Τουρκία και 1 στην Ελλάδα.

Οι προϋποθέσεις που έθεσα δίπλα από τον πίνακα πρέπει να τηρούνται αυστηρά ώστε ο αλγόριθμος του διλήματος του φυλακισμένου να τρέχει σωστά και να πάρουμε λογικά αποτελέσματα που θα μας δώσουν τα κατάλληλα συμπεράσματα.

Υπάρχουν δύο ειδών στρατηγικές που μπορούν να επιλέξουν οι δύο παίκτες στο ΔΦ. Η πρώτη είναι η φτωχή στρατηγική (Pure strategy), όπου επιλέγεις μια ενέργεια και την εκτελείς, έτσι η κάθε τιμή σε κάθε κελί του πίνακα αντιπροσωπεύει μια ενέργεια αλλά και μια φτωχή στρατηγική. Η άλλη στρατηγική είναι η μίχτη στρατηγική (Mixed strategy), όπου επιλέγει τυχαία από το σύνολο των διαθέσιμων ενεργειών, σύμφωνα με κάποια κατανομή πιθανότητας.

Από τη θεωρητική πλευρά, η επικρατέστερη ενέργεια θα ήταν αγορά στρατιωτικού εξοπλισμού δηλαδή η συνέχιση του ανταγωνισμού. Επομένως, η μόνη επικρατέστερη ισορροπία (και ισορροπία Nash 1950 [20]) είναι η κατάσταση όπου όλοι οι παίκτες

επιλέγουν την αγορά όπλων. Όμως το κοινό καλό, δηλαδή το καλό και για τους δύο, είναι η μέγιστη συνολική μακροπρόθεσμη ανταμοιβή και μεγιστοποιείται από την «Pareto αποτελεσματική» (Pareto optimality) κατανομή που επιτυγχάνεται όταν και οι δύο παίκτες συνεργαστούν [14]. Άρα, η συνολική ανταμοιβή μεγιστοποιείται μέσω συμβιβασμού.

2.5 Το Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου

Ένας από τους επιστήμονες που εντυπωσιάστηκαν από το Δίλημμα του Φυλακισμένου στα τέλη της δεκαετίας του 1970 ήταν ο Αμερικανός μαθηματικός και πολιτικός επιστήμονας Robert Axelrod [21]. Ο Axelrod δημιούργησε το Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου. Η επιλογή αυτής της κατηγορίας παιγνίου κρίνεται κατάλληλη επειδή το συγκεκριμένο παίγνιο μοντελοποιεί την εξέλιξη της ανθρώπινης συνεργασίας. Το παίγνιο αυτό δεν παίζεται μια φορά αλλά πολλές με την κάθε επανάληψη, οι παίκτες έχουν μνήμη πάνω στην ιστορία των προηγούμενων επαναλήψεων.

Παρόλο που η πιο ορθολογιστική ενέργεια για κάποιο παίγνιο $\Delta\Phi$ μίας επανάληψης είναι η επιλογή του ατομικού συμφέροντος (στρατιωτικός εξοπλισμός), όταν υπάρχει επανάληψη στο παιχνίδι, δηλαδή πολλοί γύροι, τα πράγματα αλλάζουν. Δηλαδή δεν έχουμε μια μόνο συγκεκριμένη πολιτική που είναι η καλύτερη αλλά έχουμε βέλτιστες πολιτικές έναντι των αντιπάλων του παιγνίου. Συνεπώς η επιλογή μη-συνεργασίας κάθε φορά δεν θα ήταν και η καλύτερη στρατηγική για τον λόγο ότι αν εφαρμόζεται έναντι ενός ορθολογιστικού παίχτη, τότε δεν θα πετύχουμε ποτέ συνεργασία. Ένας παίκτης θα σταματήσει να προσπαθεί να συνεργαστεί με κάποιον παίκτη που συνεχώς δε συνεργάζεται γιατί αυτό δεν είναι ορθολογιστικό. Αυτό όμως έχει ως αποτέλεσμα να φτάσουν σε αδιέξοδο αφού σε κάθε επανάληψη θα πιάνουν την ελάχιστη ανταμοιβή αφού θα είναι το κορόιδο του αντιπάλου τους. Αντίθετα, αν οι παίκτες συνεργαστούν, θα λάβουν και οι δύο μια ανταμοιβή ίση με 3, σε οποιαδήποτε επανάληψη που θα επιλέξουν και οι δύο την ενέργεια της συνεργασίας. Ανάλογα με τη στρατηγική που ακολουθεί ο αντίπαλος πρέπει οι παίκτες να μάθουν να συνεργάζονται ή να ανταγωνίζονται.

2.6 Προηγούμενη Εργασία

Σε προηγούμενες έρευνες για την επίλυση του προβλήματος της κούρσας εξοπλισμών χρησιμοποίησαν το επαναλαμβανόμενο δίλλημα φυλακισμένου(ΕΠΦ), χωρίς όμως την χρήση Ενισχυτικής Μάθησης (ΕΜ) όπου και πάλι προέκυπταν οι τέσσερις καταστάσεις που έδιναν στο τέλος της εκπαίδευσης του συστήματος τις ανάλογες πιθανότητες εμφάνισης τις κάθε κατάστασης ανάμεσα στις δύο χώρες.

Ο Majeski δήλωσε το 1984 ότι η λήξη ενός παιχνιδιού μπορεί να επάγει συνεργατική συμπεριφορά, αν σε κάθε δοσμένη επανάληψη τα μελλοντικά αναμενόμενα κέρδη συνεργασίας είναι μεγαλύτερα από τη εκμετάλλευση στη τρέχουσα επανάληψη συν τα αναμενόμενα κέρδη του αμοιβαίου ανταγωνισμού [16]. Σύμφωνα με αυτό το άρθρο ειπώθηκαν διάφορα όσο αφορά την κούρσα εξοπλισμού. Πρώτο, αν η λήξη μιας κούρσας εξοπλισμού δεν είναι ξεκάθαρη από τους συμμετέχοντες, η συνεργασία μπορεί να είναι ορθολογιστικό αποτέλεσμα μόνο αν η πιθανότητα ύπαρξης πολέμου είναι αρκετά μικρή έτσι η εκμετάλλευση θα μειώσει τις αναμενόμενες ανταμοιβές. Αν η πιθανότητα πολέμου είναι πολύ μικρή και ο αμοιβαίος συμβιβασμός ανιχνευτεί από την αρχή, τότε πράγματι δεν θα είναι λογικό να αποκαλέσουμε τέτοια κατάσταση κούρσα εξοπλισμού. Δεύτερον αν ο ένας συμμετέχοντας μειώνει σημαντικά πιθανότητα πολέμου από την δική του μεριά στη μέση μιας κούρσας εξοπλισμού, ίσως να αξίζει το ρίσκο να κινηθεί μονομερώς προς την συνεργασία έτσι ώστε να περιμένει παρόμοια απάντηση από τον αντίπαλο. Αν ένας πόλεμος δεν είναι τόσο πιθανός, τα οικονομικά οφέλη από μια συνεργασία μπορούν να αντισταθμίσουν βραχυπρόθεσμα ρίσκα ασφαλείας. Ως τρίτο αν ένα παίχτης ξεκινήσει από μόνος του συμβιβασμό και είτε ο αντίπαλος δεν ανταποκριθεί ή ανταποκριθεί θετικά και μετά εκμεταλλευτεί, τότε είναι υπεύθυνος ο αντίπαλος να ξεκινήσει από μόνος του συνεργασία. Γενικός στη κούρσα εξοπλισμού ο μη-ανταποκρινόμενος ή ο εκμεταλλευτής πρέπει το επόμενο ρίσκο και να επαναρχίσει την συνεργασία. Τέταρτο το να επιφέρουμε συνεργασία είναι λογικό όταν έθνος αντιλαμβάνεται τις μελλοντικές αναμενόμενες ανταμοιβές από τη συνεργασία να είναι μεγαλύτερες από την απώλεια από το να κάνει

εκμετάλλευση συν την αμοιβαία επίθεση. Η συνεργασία πετυχαίνεται μόνο αν ο ένας παίχτης περιμένει και ο άλλος παίχτης να αποδεχτεί τον συμβιβασμό ή τουλάχιστο όταν και οι δύο παίχτες παίζουν την ίδια στρατηγική διαφορετικά είναι δύσκολο να πετύχουν συμβιβασμό. Πέμπτο αν ένας συμμετέχοντας επιτιθόταν και εκμεταλλεύονταν τον αντίπαλο του και τώρα έχει αποφασίσει να συμβιβαστεί, τότε θα πρέπει να αναμένει να τον εκμεταλλευτεί ο αντίπαλος του για μια επανάληψη. Μια παρακινδυνευμένη κατάσταση μειώνει την εμφάνιση συμβιβασμού ακόμα και όταν γνωρίζουν και οι δύο ότι ο συμβιβασμός θα τους επιφέρει μελλοντικά καλύτερα αποτελέσματα. Τα προηγούμενα πέντε αποτελέσματα στηρίζονται στην υπόθεση ότι οι συμμετέχοντες προσπαθούν να αυξήσουν τις αναμενόμενες ανταμοιβές.

Ο Ron Smith, Martin Sola και ο Fabio Spangolo (2000), μελέτησαν την κούρσα εξοπλισμών μεταξύ Ελλάδας Τουρκίας με τη χρήση του διλήματος φυλακισμένου [15]. Συγκεκριμένα οι ανταμοιβές ειπώθηκαν να είναι αυτές του ΔΦ και επειδή δεν μπορούσαν να παρατηρήσουν ακριβώς τις ανταμοιβές αυτές δούλεψαν ως προς το να εκτιμήσουν τις στρατηγικές των δύο χωρών σε αυτό το παίγνιο. Μελέτησαν δηλαδή τις πιθανότητες που έχει η κάθε χώρα να επιλέξει υψηλή μετοχή ή χαμηλή μετοχή σε στρατιωτικό εξοπλισμό. Το παιχνίδι παιζόταν για πολλούς γύρους που ήταν τα χρόνια σε τελική ανάλυση και αυτή η πιθανότητα επιλογή υψηλής ή χαμηλής μετοχής για το επόμενο χρόνο βασιζόταν στο τρέχον χρόνο μεταξύ των στρατιωτικών εξόδων και των δύο χωρών. Η μέθοδος που χρησιμοποιήθηκε για την εκτίμηση των στρατηγικών που χρησιμοποιεί η κάθε χώρα μεταξύ της περιόδου 1958-97 ήταν η bivariate Hamilton regime-switching model. Με αυτή το μοντέλο αυτό παρατηρήθηκε ότι φαίνεται να λαμβάνει καλά τα δεδομένα, δείχνει ξεκάθαρα την διαφορά μεταξύ της υψηλής και της χαμηλής μετοχής και χωρίζει ξεκάθαρα το παιχνίδι σε τέσσερις ξεχωριστές καταστάσεις. Επίσης η απλότητα του μοντέλου τους επιτρέπει να θέσουν δύο στρατηγικές που προτείνονται από την θεωρία παιγνίων και να της εκτιμήσουν με τα δεδομένα που έχουν για να επιλέξουν ποια από τις δύο δίνει τα καλύτερα αποτελέσματα. Η μία στρατηγική είναι το 'tit-for-tat', στο οποίο κάθε κράτος κάνει ότι έκανε ο αντίπαλος του σε προηγούμενη χρονιά ενώ η άλλη υπόθεση είναι να παίζουν μια μιχτή στρατηγική όμως ανεξάρτητα χωρίς η κάθε χώρα να γνωρίζει τι έγινε σε προηγούμενο στάδιο. Όσο αφορά τη στρατηγική 'tit-for-tat' λαμβάνονται δύο σενάρια υπόψη. Το πρώτο είναι η Ελλάδα διοικεί και η Τουρκία ακολουθεί και το άλλο σενάριο έχει την Τουρκία ως αρχηγό και η Ελλάδα την ακολουθεί. Η στρατηγική 'tit-for-tat' όμως απορρίφθηκε με ένα τεστ αναλογίας πιθανότητας (likelihood ratio test) και για τα δύο σενάρια. Τώρα όσο αφορά την μιχτή στρατηγική όπου η δύο χώρες επιλέγουν ενέργεια ανεξάρτητα με το τι κάνει ο αντίπαλος ή με το τι έγινε πριν έγινε αποδεκτή από το τεστ αναλογίας πιθανότητας (likelihood ratio test). Επίσης παρατηρήθηκε ότι οι εκτιμήσεις σε 'εξωτερικό τομέα', δράσης-αντίδρασης για μετοχές σε στρατιωτικές δαπάνες δεν είναι συνεπής, αντίθετα με τις εκτιμήσεις σε 'εσωτερικό' τομέα που είναι συνεπείς και οι οποίες με την σειρά τους δίνουν έμφαση στην γραφειοκρατία και την πολιτική αδράνεια.

Οι Vassiliades, Christodoulou, Cleanthous (2011) ερευνούν τη ενισχυτική μάθηση πολλαπλών πρακτόρων σε παίγνια γενικού αθροίσματος με spiking και non-spiking πράκτορες στο επαναλαμβανόμενο δίλλημα του φυλακισμένου για να εξερευνήσουν τις καταστάσεις που χρειάζονται για να ενισχύσουν το αποτέλεσμα του συμβιβασμού [17]. Η spiking προσέγγιση περιλαμβάνει νευρωνικά δίκτυα με νευρώνες Leaky Integrate and Fire εκπαιδευόμενους με δύο διαφορετικούς αλγορίθμους μάθησης: Ενίσχυση της διαβίβασης της σύναψης και ενίσχυση του spike-timing depended plasticity με ίχνος εκλεξιμότητας (STDP- η πλαστικότητα εξαρτάται από το χρόνο άφιξης των δυναμικών ενεργείας). Η non-spiking γίνεται με αλγορίθμους Q και SARSA. Τα αποτελέσματα του συμβιβασμού ενισχύονται σύμφωνα με τα εσωτερικά μεταδιδόμενα σήματα ενίσχυσης και συνδυασμό ψηλού ρυθμού μάθησης και χαμηλού ρυθμού έκπτωσης με μια κατάλληλη εξερεύνηση στη περίπτωση του non-spiking πράκτορα. Ενώ στη περίπτωση του spiking πράκτορα τα αποτελέσματα του συμβιβασμού ενισχύθηκαν σύμφωνα με ένα μακρύτερο ίχνος εκλεξιμότητας που έθεσαν. Επιπρόσθετα έδειξαν ότι και οι δύο προσεγγίσεις έχουν παρόμοια συμπεριφορά με αποτέλεσμα να μπορούν να χρησιμοποιηθούν και τα δύο σε αλληλεπίδρασης πολλαπλών πρακτόρων δίνοντας τα ίδια αποτελέσματα.

Κεφάλαιο 3

Περιγραφή συστήματος

3.1 Εισαγωγή	24
3.2 Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων	26
3.3 Συστατικά Συστήματος	31

3.1 Εισαγωγή

Χρησιμοποιώντας την γλώσσα προγραμματισμού Java υλοποίησα την διπλωματική μου εργασία, μια γλώσσα που περιέχει αντικειμενοστρέφεια και είναι απλή, αποδοτική και ευέλικτη να τρέξει προγράμματα σε διαφορετικά λειτουργικά συστήματα όπως Windows, Linux, Unix και Macintosh.

Στα πλαίσια αυτής της διπλωματικής θα δημιουργήσουμε ένα σύστημα στο οποίο οι πράκτορες θα εκπαιδεύονται με Ενισχυτική Μάθηση, βασιζόμενοι στο παίγνιο του Επαναλαμβανόμενου Διλήμματος του Φυλακισμένου. Με την χρήση της Ενισχυτικής Μάθησης θα αξιολογηθεί ο κάθε πράκτορας ξεχωριστά. Θα αναπτύξουμε ένα κώδικα σχετικά απλό αλλά και αποδοτικό με τρόπο που να μπορεί να επεκταθεί με διάφορους τρόπους. Δουλειά αυτού του συστήματος είναι η δημιουργία ενός γενικού συστήματος που να υποστηρίζει την ανάπτυξη πρακτόρων και την αξιολόγηση των ενεργειών τους κατά την διάρκεια του χρόνου εκπαιδεύοντας το με EM.

Επιπρόσθετα το σύστημα που χρειαζόμαστε πρέπει να είναι σε θέση να χειρίζεται αλληλεπιδράσεις μεταξύ πολλαπλών πρακτόρων (όχι μόνο δύο) στο παίγνιο, αφού αργότερα θα μπορούσε να υλοποιηθούν και παίγνια n παικτών.

Επίσης κύρια βάση του συστήματος που θα δημιουργηθεί όπως είπαμε είναι το «Επαναλαμβανόμενο Δίλημα του Φυλακισμένου», έτσι θα δημιουργηθούν τέσσερις διαφορετικές καταστάσεις, που η κάθε κατάσταση θα πρέπει να αξιολογείται για κάθε πράκτορα ξεχωριστά σύμφωνα με τη δράση που θα επιλέξει. Επομένως μέσα στο σύστημα θα υπάρχουν και οι ανάλογες ανταμοιβές σύμφωνα με τις τέσσερις καταστάσεις και των ενεργειών που επιλέγει κάθε φορά ο πράκτορας. Κύριο σημείο του συστήματος είναι η υλοποίηση του αλγορίθμου μάθησης q-learning, ο οποίος θα δίνει την αξιολόγηση του κάθε πράκτορα ξεχωριστά σε κάθε γύρο. Επίσης στο σύστημα θα προστεθεί η ιδιότητα ε-greedy με την οποία θα επιλέγεται η ενέργεια του κάθε πράκτορα για να πετύχουμε περισσότερη εξερεύνηση (exploration).

Το σύστημα καλό θα ήταν να σχεδιαστεί με τρόπο που να προσαρμόζεται εύκολα σε νέα δεδομένα. Επομένως αν θέλει κάποιος να το επεκτείνει προσθέτοντας του νέα περιβάλλοντα, παίγνια, πράκτορες, παραμέτρους, πολιτικές, μετρικές), να μην χρειάζεται να ξέρει πως λειτουργεί το σύστημα, αλλά να πρέπει μόνο να επεκτείνει συγκεκριμένες κλάσεις, γράφοντας τον δικό του κώδικα, χωρίς να χρειάζεται να αλλάξει τα κύρια συστατικά του συστήματος.

3.2 Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων

Ως πρώτο στάδιο αποφασίσαμε να χωρίσουμε το σύστημα μας σε δύο μέρη για να αποφύγουμε τον κίνδυνο ο ένας παίχτης εκ των δύο να θεωρηθεί μέρος του περιβάλλοντος και να αλλάζει δυναμικά. Ως αρχική κατάσταση βάλαμε το σύστημα να διαλέξει τυχαία μια από τις τέσσερις καταστάσεις, τις οποίες εξηγήσαμε σε προηγούμενα κεφάλαια.

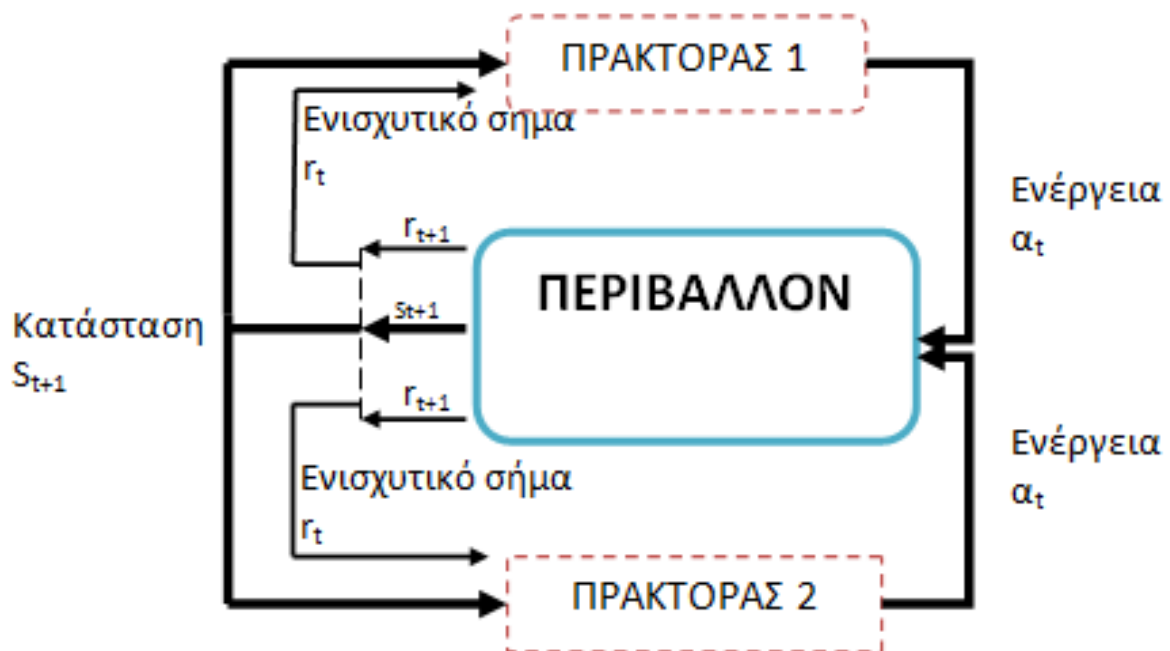
Χωρίσαμε τον αλγόριθμο σε δύο φάσεις. Σε αυτή την φάση συναντούμε την αρχικοποίηση της κατάσταση του περιβάλλοντος που πετυχαίνεται με την τυχαία επιλογή από τις τέσσερις υπάρχοντες καταστάσεις. Επίσης έχουμε την αρχικοποίηση των τιμών του πίνακα αξιών για κάθε πράκτορα καθώς και κάποιες άλλες αρχικοποιήσεις παραμέτρων όπως τον ρυθμό μάθησης, ρυθμό έκπτωσης και ϵ (πιθανότητα που χρησιμοποιείται για να επιτρέψει τους πράκτορες να κάνουν εξερεύνηση).

Στη δεύτερη φάση έχουμε πολλές επαναλήψεις (run) από πολλούς γύρους, όπου για κάθε γύρο (round) έχουμε την διαδικασία εκπαίδευσης. Κάθε πράκτορας επιλέγει την ενέργειά του με βάση την κατάσταση του περιβάλλοντος σύμφωνα με μια πιθανότητα ϵ που θα εξηγήσουμε παρακάτω. Ακολούθως, το περιβάλλον συλλέγει τις ενέργειες που επέλεξαν οι δύο πράκτορες και με αυτές διαμορφώνει τη νέα του κατάσταση. Ανάλογα με την ενέργεια που θα επιλέξουν και σύμφωνα με την κατάσταση στην οποία βρίσκονται ο κάθε πράκτορας θα ανανεώνει τη συνάρτηση αξίας δηλαδή, μια Q-value τιμή που θα υπολογίζεται από την πράξη που μας δίνεται από τον αλγόριθμο Q-learning και πριν το τέλος κάθε γύρου η τρέχον κατάσταση θα γίνει ίση με την επόμενη κατάσταση.

Τέλος, μετά από κάθε επανάληψη συλλέγονται δεδομένα για το πόσες φορές εμφανίστηκε η κάθε κατάσταση (πιθανότητα εμφάνισης) σε κάθε γύρο για να έχουμε στατιστικά δεδομένα τα οποία θα δούμε και θα συζητήσουμε σε επόμενα κεφάλαια. Μια επανάληψη

(run) καλό θα ήταν να αποτελείται από πολλούς γύρους(round) για να λαμβάνουμε έτσι την πιθανότητα εμφάνισης τις κάθε καταστάσεις ώστε να δούμε αν πετυχαίνεται ο σκοπός μας, που είναι να έχουμε μεγαλύτερη πιθανότητας εμφάνισης της κατάστασης συνεργασίας μεταξύ των πρακτόρων. Με αυτές τις πιθανότητες θα στηρίξουμε και τα συμπεράσματά μας.

Στο Σχήμα 3.1 μπορούμε να δούμε τον τρόπο αλληλεπίδρασης 2 πρακτόρων στο σύστημά μας και το Σχήμα 3.2 μπορούμε να δούμε τον ψευδοκώδικα του αλγορίθμου ενισχυτικής μάθησης πολλαπλών πρακτόρων, που περιγράφηκε πιο πάνω.



Σχήμα 3.1: Αλληλεπίδραση Περιβάλλοντος με τους 2 Πράκτορες

Κάθε πράκτορας σε μια δεδομένη χρονική στιγμή παρατηρεί την κατάσταση του περιβάλλοντος καθώς και την επιβράβευση/τιμωρία (ενισχυτικό σήμα) που λαμβάνει και επιλέγει μια ενέργεια ανάλογα με την πολιτική που ακολουθεί. Οι ενέργειες δίνουν τη νέα κατάσταση του περιβάλλοντος και έτσι στην επόμενη χρονική στιγμή το περιβάλλον στέλλει τη νέα του κατάσταση στους πράκτορες μαζί με τα νέα ενισχυτικά σήματα που κρίνουν τις ενέργειες τους.

```
Agent Turkey=new Agent()  
Agent Greece=new Agent() //Δημιουργία 2 πρακτόρων Τουρκίας- Ελλάδας  
numberOfRuns=10//Αριθμός Επαναλήψεως  
numberOfRounds=500//αριθμός γύρων σε κάθε τρέξιμο
```

```
action=[action1,action2]// action1-ενέργεια που επιλέγει η  
Τουρκία,  
//action2-ενέργεια που επιλέγει η Ελλάδα.
```

```
Reward=[{3,3}, //<NotBuild,NotBuild> ενισχυτικό σήμα  
{1,4},//<NotBuild,Build>  
{4,1},//<Build,NotBuild>  
{2,2}}]//< Build,Build >
```

Αυτές είναι ενδεικτικές τιμές που χρησιμοποιήσαμε για να ελέγξουμε την ορθότητα του συστήματος. Μεταγενέστερα χρησιμοποιήσαμε τιμές από ιστορικά γεγονότα.

Φάση 1:Αρχικοποίηση και εύρεση της αρχικής κατάστασης

A)

```
Turkey.initialize();  
Greece.initialize();
```

Οι q-values τιμές του κάθε πράκτορα αρχικά αρχικοποιούνται με την πράξη $r_{max} / (1 - \text{GAMMA})$ που έδινε περισσότερη εξερεύνηση όμως σε τελικό στάδιο αρχικοποιήσαμε με 0.

B)

Για κάθε αριθμό τρεξίματος (Run):

Για τη αρχική κατάσταση του περιβάλλοντος επιλέγουμε μια κατάσταση τυχαία από τις 4 υπάρχουσες.

```
currentstate = rand.nextInt(4);
```

Γ)

Αυτή η κατάσταση στέλνεται σε μια μέθοδο *train()* όπου γίνεται η εκπαίδευση του συστήματος με το πέρασμα κάποιον γύρων (Round).

```
train(currentstate);
```

Φάση 2:Εκπαίδευση του συστήματος

Για κάθε γύρο:

1)Ο κάθε πράκτορας βασιζόμενος στη τρέχον κατάσταση επιλέγει ενέργεια.

```
action1=Turkey.chooseAction(state, epsilon);
```

```
action2=Greece.chooseAction(state, epsilon);
```

2)Λαμβάνουν το ενισχυτικό σήμα σύμφωνα με την ενέργεια που επέλεξαν στη δεδομένη κατάσταση.

```
rewardTurkey=Turkey.setReward(action1,state);
```

```
rewardGreece=Greece.setReward(action2,state);
```

3)Παίρνουμε την επόμενη κατάσταση σύμφωνα με τις ενέργειες που επέλεξαν τα δύο κράτη.

```
nextState=getState(action1,action2);
```

4)Τέλος ενημερώνουμε τις q-values τιμές του κάθε πράκτορα

```
Turkey.updateValues(state, action1, nextState, rewardTurkey);
```

```
Greece.updateValues(state, action2, nextState, rewardGreece);
```

5) Η τρέχον κατάσταση να γίνει η επόμενη κατάσταση

```
state=nextState;
```

6)Υπολογισμοί για τις πιθανότητες εμφάνισης της κάθε κατάστασης στο αντίστοιχο γύρο.

Προσομοίωση

ΦΑΣΗ 1

ΦΑΣΗ 2

Αποτελέσματα σε πιθανότητες εμφάνισης(0-1)

Τελικό ποσοστό εμφάνισης της κάθε κατάστασης(%)

Σχήμα 3.2: Ενισχυτικής Μάθησης Πολλαπλών Πρακτόρων (Ψευδοκώδικας)

Ο αλγόριθμος που χρησιμοποιήσαμε, χωρίζεται σε 2 φάσεις. Κατά τη πρώτη φάση οι πράκτορες δημιουργούνται και αρχικοποιούν τις q-values τιμές με την τιμή 0. Μετά για κάθε τρέξιμο επιλέγει τυχαία μια κατάσταση από τις 4 υπάρχουσες και την στέλνει σαν παράμετρο στη μέθοδο train για να γίνει η εκπαίδευση. Κατά τη δεύτερη φάση , για κάθε γύρο ενός τρεξίματος, ο κάθε πράκτορας παρατηρεί το περιβάλλον και επιλέγει μια ενέργεια σύμφωνα με μια πιθανότητα epsilon (μηχανισμός εξερεύνησης). Το περιβάλλον δίνει τη νέα του κατάσταση και τις ανταμοιβές των πρακτόρων και ο κάθε πράκτορας ενημερώνει τις q-values τιμές του, ανάλογα με την ανταμοιβή που έλαβε. Στο τέλος το σύστημα συλλέγει κάποια στατιστικά δεδομένα που είναι η πιθανότητα εμφάνισης τις κάθε κατάστασεις που θα χρησιμεύσουν στην πορεία για να βγάλουμε τα ανάλογα συμπεράσματα.

3.3 Συστατικά Συστήματος

Το σύστημα μου περιέχει 2 χαρακτηριστικά τα οποία είναι τα παρακάτω:

A) Πράκτορας(Agent)

B) Προσομοίωση(Simulation)

Η κλάση Agent είναι η σημαντικότερη κλάση από την έννοια ότι μέσα στην κλάση αυτή βρίσκονται οι περισσότερες λειτουργίες του συστήματος. Υλοποίησα ξεχωριστή κλάση για τον πράκτορα ούτως ώστε να μπορούμε να δημιουργήσουμε όσους πράκτορες θέλουμε για υλοποίηση σε μελλοντική εργασία. Σε αυτή τη διπλωματική εργασία δημιούργησα αυτή την κλάση ώστε να μπορώ να δημιουργήσω 2 διαφορετικούς πράκτορες. Ο ένας είναι ο πράκτορας που θα αντιπροσωπεύει την Τουρκία ενώ ο άλλος πράκτορας θα αντιπροσωπεύει την Ελλάδα.

Τώρα όμως ας αφοσιωθούμε στις μεθόδους (λειτουργίες) που δημιουργήθηκαν σε αυτή την κλάση-πράκτορας. Επίσης πολύ σημαντικό είναι το ενισχυτικό σήμα που παίρνει ο κάθε πράκτορας. Έτσι μέσα σε αυτή την κλάση έχουμε ένα πίνακα στον οποίο αποθηκεύουμε τις ανταμοιβές του κάθε πράκτορα στη δεδομένη κατάσταση με τη συγκεκριμένη δράση. Να αναφέρουμε ότι οι τιμές αυτές είναι και οι τιμές που δηλώσαμε και στο δίλλημα του φυλακισμένου και οι τιμές που χρησιμοποιήσαμε επιλέχτηκα λαμβάνοντας υπόψη κάποια ιστορικά δεδομένα. Επιπρόσθετα έχουμε ένα πίνακα μέσα στον οποίο αποθηκεύονται η q-value τιμές για τον συγκεκριμένο πράκτορα στη δεδομένη κατάσταση που επέλεξε τη συγκεκριμένη ενέργεια. Αυτός ο πίνακας είναι το κρίσιμο σημείο του προγράμματος αφού χρησιμοποιείται στο τέλος για να βγάλουμε τις πιθανότητες εμφάνισης της κάθε κατάστασης για να ελέγξουμε ποια κατάσταση υπερτερεί τον άλλο. Με πιο απλά λόγια θέλουμε να δούμε αν το σύστημα μας μαθαίνει το συμβιβασμό έχοντας στη κατάσταση συμβιβασμού την πιο μεγάλη πιθανότητα εμφάνισης.

Ο πίνακας που αποθηκεύει τις q-value τιμές πρέπει να αρχικοποιηθεί. Επομένως αρχικοποιήσαμε τον πίνακα αυτό με τιμή ίση με 0.

Ακόμα μια μέθοδος που βρίσκεται στη κλάση Agent είναι η `chooseAction(state, epsilon)`, η οποία βασιζόμενη στη δεδομένη κατάσταση που παίρνει ως παράμετρο επιστρέφει την ανάλογη ενέργεια στο κάθε πράκτορα. Η ενέργεια που θα επιστρέψει όμως επιλέγεται τυχαία με πιθανότητα `epsilon`. Συγκεκριμένα αν η τυχαία τιμή είναι μικρότερη από την τιμή του `epsilon` τότε θα επιστραφεί τυχαία ενέργεια. Αντίθετα αν η τυχαία τιμή είναι μεγαλύτερη του `epsilon` θα επιστραφεί `greedy` ενέργεια. Η `greedy` ενέργεια είναι η μέγιστη q-value τιμή στη κατάσταση που δίνεται σαν παράμετρος σε μια άλλη μέθοδο τη `greedyAction(state)`, που καλείται από την `chooseAction(state, epsilon)`, όταν η τυχαία τιμή είναι μεγαλύτερη από την τιμή `epsilon` όπως αναφέραμε. Αρχικά το `epsilon` παίρνει τιμή 1 και με το πέρασμα κάθε γύρου μειώνεται, με αυτό τον τρόπο δύνουμε τη δυνατότητα εξερεύνησης στους δύο παίκτες.

Η μέθοδος που επιστρέφει το ενισχυτικό σήμα σε κάθε πράκτορα βρίσκεται και αυτή στην κλάση Agent. Παίρνει παραμέτρους την ενέργεια και την κατάσταση που βρίσκεται ο κάθε πράκτορας και του επιστρέφει την ανάλογη ανταμοιβή σύμφωνα με τον πίνακα ανταμοιβών που αναφέρθηκε προηγουμένως.

Υπάρχει επίσης η μέθοδος `updateValues` όπου ανανεώνονται η τιμές του πίνακα q σύμφωνα με τη πράξη του αλγόριθμο q-learning και τις ανάλογες παραμέτρους που είναι , η τρέχον κατάσταση, η ενέργεια, η επόμενη κατάσταση και η άμεση ανταμοιβή. Εδώ να αναφέρουμε ότι χρησιμοποιείται ο ρυθμός έκπτωσης. Ο οποίος έχει αρκετά μεγάλη τιμή αφού θέλουμε να πάρουμε την μεγαλύτερη ολική q-value τιμή μακροπρόθεσμα και όχι άμεσα. Επιπρόσθετα έχουμε και την τιμή του ρυθμού μάθησης μέσα σε αυτή την πράξη

ούτως ώστε να γίνεται μάθηση του συστήματος, έτσι και αυτή η τιμή είναι ψηλή για να πετύχει το σκοπό της. Ακόμη σε αυτή τη μέθοδο καλούμε μια άλλη μέθοδο την *maximum*, που θα μας επιστρέψει την μέγιστη *q-value* τιμή σύμφωνα με την επόμενη κατάσταση και τις πιθανές ενέργειες.

Τώρα ας αναφερθούμε στην άλλη μας κλάση την *simulation* η οποία δημιουργεί τους 2 πράκτορες Ελλάδα και Τουρκία. Δηλαδή δημιουργεί δύο στιγμιότυπα της κλάσης *Agent*, ώστε να μπορεί να χειριστεί τις λειτουργίες αυτής της κλάσης για κάθε πράκτορα ξεχωριστά. Και αυτή η κλάση είναι πολύ σημαντική αφού εδώ καλούνται όλες οι μέθοδοι της *Agent* για να γίνει η εκπαίδευση του συστήματος. Αφού δημιουργήσουμε τα στιγμιότυπα για Ελλάδα και Τουρκία δηλώνουμε επίσης τον αριθμό των επαναλήψεων και τον αριθμό των γύρων σε κάθε επανάληψη. Επίσης δημιουργούμε τέσσερις πίνακες ένα για κάθε κατάσταση, στους οποίους το σύστημα θα αποθηκεύει πόσες φορές εμφανίστηκε η κάθε κατάσταση σε κάθε γύρο. Στο τέλος θα χρησιμοποιηθούν αυτή οι πίνακες για να βγάλουμε μέσους όρους για κάθε κατάσταση από όλους τους γύρους μετά από όλες τις επαναλήψεις. Αυτοί οι μέσοι όροι θα είναι και οι πιθανότητες εμφάνισης της κάθε κατάστασης. Επίσης στο τέλος όλων των γύρων υπολογίζεται και το ποσοστό εμφάνισης της κάθε κατάστασης.

Ως πρώτο βήμα για να ξεκινήσει ο αλγόριθμος είμαστε στη μέθοδο *simulate* αρχικοποιούμε τους πίνακες αξιών *q* με 0 για την Ελλάδα και μια για την Τουρκία. Στην συνέχεια για κάθε επανάληψη θα υπολογίζεται τυχαία η αρχική κατάσταση του περιβάλλοντος σύμφωνα με τις υπάρχοντες τέσσερις καταστάσεις. Αυτή η αρχική κατάσταση του περιβάλλοντος θα περνά ως παράμετρος σε μια μέθοδο που θα εκπαιδεύει τον αλγόριθμο, την *train(state)*.

Τώρα σε αυτή την μέθοδο για κάθε γύρο της δεδομένης επανάληψης, οι δύο πράκτορες Τουρκία και Ελλάδα θα επιλέξουν δράση καλώντας τη μέθοδο *chooseAction* της κλάσης *Agent* ξεχωριστά. Αυτή η επιλογή θα γίνει σύμφωνα με αυτά που προαναφέρθηκαν στην

κλάση Agent λαμβάνοντας υπόψη την κατάσταση στην οποία βρίσκονται. Παίρνοντας τώρα τις δύο ενέργειες, επιστρέφουμε την ανάλογη ανταμοιβή για τον κάθε πράκτορα ξεχωριστά καλώντας την μέθοδο `setReward` της κλάσης Agent. Ενώ στη συνέχεια παίρνοντας και πάλι τις δύο ενέργειες και καλούμε μια μέθοδο που θα μας επιστρέψει την επόμενη κατάσταση την `getState` που είναι της κλάσης `simulation`. Μετά ενημερώνουμε τις τιμές της συνάρτησης αξίας (δηλαδή το πίνακα `q`) καλώντας την `updateValues` για κάθε πράκτορα ξεχωριστά και πριν το τέλος του κάθε γύρου η τρέχον κατάσταση γίνεται ίση με την επόμενη κατάσταση. Μέσα σε κάθε γύρο επίσης υπολογίζουμε πότε εμφανίζεται η κάθε κατάσταση και ενημερώνουμε τον αντίστοιχο πίνακα της κατάστασης. Μόλις τελειώσουν όλες οι επαναλήψεις εμφανίζουμε τους μέσους όρους για την κάθε κατάσταση καθώς και το ποσοστό εμφάνισης.

Κεφάλαιο 4

Αντιστοίχιση Κούρσας εξοπλισμών στο «Επαναλαμβανόμενο Δίλλημα Φυλακισμένου» και Αποτελέσματα

4.1 Πρώτη Φάση	37
4.2 Δεύτερη Φάση	44
4.3 Τρίτη Φάση	52
4.4 Τέταρτη Φάση	60
4.5 Συνένωση Προηγούμενων Φάσεων	68

Αρχικά θα ήθελα να αναφέρω ότι σε αυτό το κεφάλαιο της αντιστοίχισης ανταμοιβών βασίστηκα στα ιστορικά δεδομένα που πήρα από το άρθρο “The Prisoner’s Dilemma and Regime-Switching in the Greek-Turkish Arm Race” των Smith, Sola και Spangolo όσο αφορά τις δαπάνες που έγιναν και από τις δύο χώρες σε στρατιωτικό εξοπλισμό.

Κύριος σκοπός αυτής της διπλωματικής εργασίας είναι η επίλυση της κούρσας εξοπλισμών μεταξύ Ελλάδας και Τουρκίας. Θα εξηγήσουμε πως προκύπτουν οι αμοιβές σε ένα πίνακα αμοιβών στο «επαναλαμβανόμενο δίλλημα του φυλακισμένου» ανάλογα με διάφορες φάσεις που χαρακτηρίζουν την κάθε χώρα και θα αναλύσουμε την κάθε φάση ξεχωριστά.

Σύμφωνα με πρόσφατα δεδομένα η Ελλάδα σπαταλάει 4,37% **ΑΕΠ** (σύνολο όλων των προϊόντων και αγαθών που παράγει μια οικονομία, εκφρασμένο σε χρηματικές μονάδες) το ελάχιστο σε στρατιωτικό εξοπλισμό και 6,53% του **ΑΕΠ** του το μέγιστο. Αντίθετα με την Τουρκία που ξοδεύει το ελάχιστο 3,86% και 5,30% το μέγιστο. Είναι προφανές ότι η Ελλάδα ξοδεύει περισσότερο από την Τουρκία σε γενικό βαθμό και αυτό συμβαίνει επειδή η Ελλάδα έχει λιγότερο πληθυσμό από την Τουρκία. Έτσι ο μόνος τρόπος να σταθούν αντάξιοι αντίπαλοι είναι η Ελλάδα να ξοδεύει περισσότερα ούτως ώστε να σταθεί αντάξια στην Τουρκία.

Τώρα όσο αφορά τον πίνακα αμοιβών, αρχικά πρέπει να ορίσουμε όρια τιμών. Οι τιμές θα βασίζονται στα έξοδα σε στρατιωτικό εξοπλισμό δηλαδή όταν και οι δύο χώρες αποφασίσουν να μην αγοράσουν εξοπλισμό, δεν θα έχουν έξοδα εξοπλισμού έτσι θα πάρουν την μέγιστη ανταμοιβή του πίνακα τιμών.

Αυτή η μέγιστη τιμή θα είναι ανάλογη τον πιο πάνω ποσοστών, άρα θα είναι ανάλογη τον μέγιστων εξόδων που θα γλιτώσουν όταν δεν θα αγοράσουν άλλο εξοπλισμό και της χρονολογικής κατάστασης στην οποία βρίσκεται μια χώρα. Αυτό που εννοώ με την χρονολογική κατάσταση είναι ότι, σε ορισμένες περιόδους της προϊστορίας αυτές οι χώρες σπαταλούσαν πολύ περισσότερα σε εξοπλισμό έναντι σε άλλες περιόδους. Επομένως θα ορίσουμε 65 την μέγιστη τιμή που μπορούν να πάρουν οι δύο χώρες μόνο αν επιλέξουν και οι δύο να μην αγοράσουν εξοπλισμό αλλά βρίσκονταν και σε περίοδο, που ξοδευόταν μεγάλος όγκος χρημάτων σε στρατιωτικό εξοπλισμό. Ανάλογα η μια χώρα έκτον δυο θα πάρει την ανταμοιβή 65 λόγω περισσότερης σπατάλης (Ελλάδα στην περίπτωση μας), ενώ η άλλη θα πάρει ελαφρά μικρότερη ανταμοιβή.

Σύμφωνα με την ελάχιστη τιμή του πίνακα τιμών(«τιμωρία»), θα είναι ανάλογη των μέγιστων εξόδων που θα σπαταλήσουν για να αγοράσουν εξοπλισμό αλλά και της περιόδου στην οποία γίνεται μεγάλη δαπάνη χρημάτων σε στρατιωτικό εξοπλισμό. Άρα η ελάχιστη τιμή που μπορούν να πάρουν είναι η -65(για τιμωρίες θα δίνονται αρνητικές τιμές), και θα τη πάρει μόνο εκείνος που θα αποφασίσει να μην αγοράσει εξοπλισμό ενώ ο ανταγωνιστής του θα αγοράσει, σε κρίσιμη περίοδο (μεγάλης δαπάνης χρημάτων). Ως συμπέρασμα τα όρια του πίνακα τιμών θα είναι πλέον [-65,65].

4.1 Πρώτη φάση

Ως πρώτη φάση έχουμε την περίοδο 1974-83 όπου γίνεται μεγάλη δαπάνη χρημάτων και από τις δύο χώρες, όμως σε γενικό βαθμό η Ελλάδα ξοδεύει 1,23% περισσότερο από την Τουρκία λόγω του μικρότερου πληθυσμού της. Μέσα από τις διάφορες δοκιμές για την εύρεση των κατάλληλων τιμών πρόσεξα ότι βάζοντας μεγαλύτερη διαφορά τιμής μεταξύ ανταμοιβών και τιμωριών μεταξύ Ελλάδας και Τουρκίας είχαμε περισσότερη δυσκολία του να φτάσουν αυτές οι δύο χώρες σε συμβιβασμό (θα φανεί και αργότερα στη γραφική). Επομένως παρόλο που βρισκόμαστε σε περίοδο που και οι δύο ξοδεύουν πληθώρα χρημάτων σε στρατιωτικό εξοπλισμό η Ελλάδα σε γενικό βαθμό έχει ένα ξεκάθαρο προβάδισμα ως πιο σπάταλη από την Τουρκία. Έτσι αυτή η διαφορά τιμής μεταξύ των χωρών όντως πρέπει να είναι αρκετά μεγάλη για να ξεχωρίζει ότι η Ελλάδα ξοδεύει περισσότερο.

Σύμφωνα με αυτά τα γεγονότα ακολουθεί ο πίνακας με τις θέσεις των μερών και μια πιθανή αντιστοίχιση σε καταστάσεις ενός παιγνίου του «δίλλημα του φυλακισμένου». Η Τουρκία παίρνει την αμοιβή με τα κόκκινα γράμματα ενώ η Ελλάδα την αμοιβή με τα μπλε.

Πινάκας 1: Αντιστοίχιση αμοιβών με ρυθμό έκπτωσης 0,9 και ρυθμό μάθησης 0,9 και οι δύο χώρες εξοπλίζονται σε μεγάλους ρυθμούς. Οι τιμές είναι αντικατοπτρίστηκες για την περίοδο 1974-83.

Τουρκία	Ελλάδα	
	Συμβιβασμός	Ελλάδα εξοπλίζεται
	• Να προβούν και οι δύο χώρες σε μια συμφωνία για να κρατηθεί ειρήνη. • Να βρεθεί ένα ισοζύγιο των εξοπλισμών μεταξύ των δύο χωρών.	• Η Ελλάδα δαπανεί μεγάλο μέρος των χρημάτων της σε στρατιωτικό εξοπλισμό(περισσότερο από την Τουρκία), έτσι χάνει μέρος της οικονομίας. • Η Τουρκία δεν εξοπλίζεται περεταίρω. Άρα πλήγμα προς τη στρατιωτική της άμυνα, διατηρεί όμως το μέρος της οικονομίας της που υπό άλλες συνθήκες θα δινόταν για αγορά εξοπλισμού.
	NN1(40,65)	NB1(-40,0)
	Τουρκία εξοπλίζεται	Επίθεση
	• Η Τουρκία δαπανεί μεγάλο μέρος των χρημάτων(λιγότερο από την Ελλάδα) της σε στρατιωτικό εξοπλισμό, έτσι χάνει μέρος της οικονομίας. • Η Ελλάδα δεν εξοπλίζεται περεταίρω. Άρα πλήγμα προς τη στρατιωτική της άμυνα, διατηρεί το μέρος της οικονομίας της όμως που υπό άλλες συνθήκες θα δινόταν για αγορά εξοπλισμού.	• Τουρκία και Ελλάδα εξοπλίζονται ασύστολα για να υπερισχύσουν, έτσι προκαλούν πλήγμα στην οικονομία τους(Η Ελλάδα λίγο μεγαλύτερο πλήγμα-μεγαλύτερη δαπάνη). • Ύπαρξη διαμάχης μεταξύ τους λόγω της υπόνοιας ύπαρξης πολέμου.
	BN1(25,-65)	BB1(-10,-35)

Όπως παρατηρείται από των πίνακα αμοιβών έχουμε ορίσει μια διαφορά +25 μεταξύ των δύο χωρών δηλαδή η Ελλάδα παίρνει +25 μεγαλύτερη αμοιβή ή τιμωρία από την Τουρκία όπως αναφέραμε πιο πάνω. Ας μιλήσουμε ξεχωριστά για την κάθε περίπτωση. Όταν είμαστε στη περίπτωση του συμβιβασμού η Ελλάδα θα πάρει μια μεγαλύτερη ανταμοιβή (65) από την Τουρκία (40) και μάλιστα θα πάρει την μέγιστη ανταμοιβή επειδή βρισκόμαστε σε περίοδο «αιχμής» και οι δύο χώρες σπαταλούν ασύστολα στρατιωτικό εξοπλισμό με την Ελλάδα να ξοδεύει γενικός περισσότερα, έτσι με το να φτάσουν σε ένα συμβιβασμό η Ελλάδα θα γλιτώσει περισσότερο μέρος των χρημάτων της από την Τουρκία επομένως έχουμε αυτή την διαφορά (+25) στις ανταμοιβές των δύο χωρών.

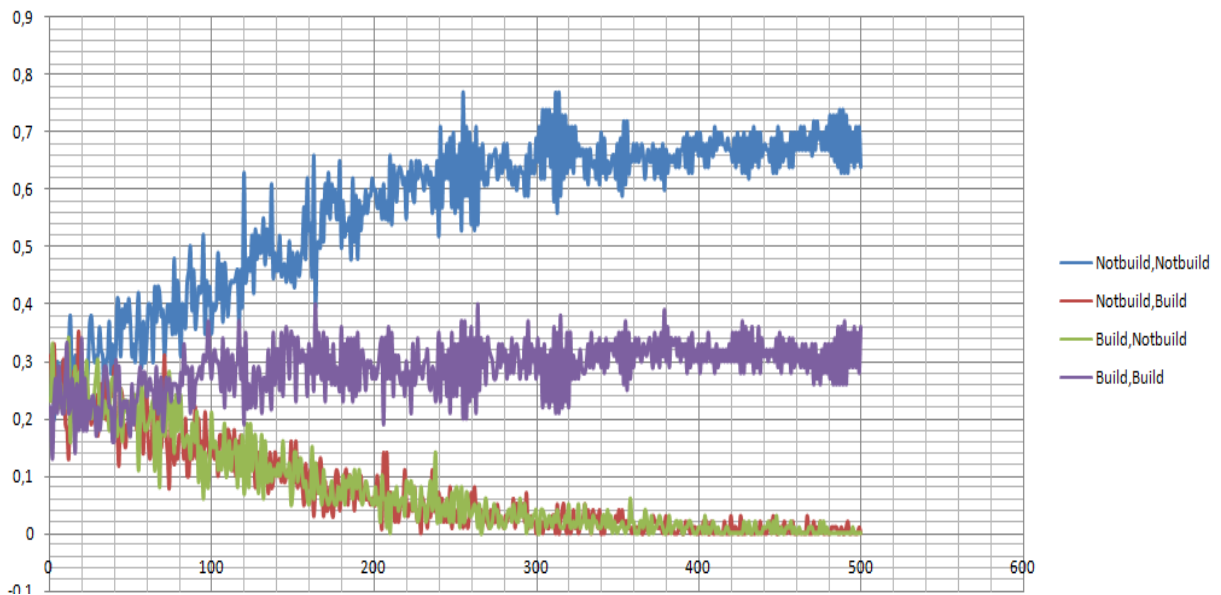
Όταν είμαστε στη περίπτωση που η Ελλάδα ξοδεύει τότε η Ελλάδα θα πάρει την μικρότερη επιβράβευση 0 γιατί μπορεί να υπερτερεί σε στρατιωτικό εξοπλισμό από την Τουρκία με αυτά που ξοδεύει όμως με αυτό τον τρόπο πλήττει την οικονομία της ανησυχητικά. Η Τουρκία με την σειρά της παίρνει μια τιμωρία (-40) αφού γίνεται το κορόιδο σε αυτή την περίπτωση της Ελλάδας με τον να μην ξοδέψει πολλά σε εξοπλισμό.

Τώρα στην περίπτωση που η Τουρκία εξοπλίζεται θα πάρει μεγαλύτερη ανταμοιβή (25) από εκείνη που πήρε η Ελλάδα όταν εξοπλίστηκε αλλά πάλι είναι σχετικά μικρή γιατί βρισκόμαστε σε περίοδο μεγάλης σπατάλης και έχουν κακές επιπτώσεις στην οικονομία τους. Επίσης η ανταμοιβή της Τουρκίας είναι +25 σε σχέση με της Ελλάδας στη προηγούμενη περίπτωση επειδή δεν ξοδεύει σε τόσο βαθμό όσο η Ελλάδα άρα θα έχει μικρότερο πλήγμα η Τουρκία στην οικονομία της άρα θα πάρει μεγαλύτερη ανταμοιβή. Τώρα όσο αφορά την Ελλάδα σε αυτή την περίπτωση του εξοπλισμού της Τουρκίας, η Ελλάδα γίνεται κορόιδο και θα πάρει την μέγιστη τιμωρία (-65) αφού δεν εξοπλίζεται στρατιωτικά σε μια περίοδο που και οι 2 χώρες εξοπλίζονται σε μεγάλο βαθμό.

Τώρα στην τελευταία περίπτωση έχουμε την επίθεση, όπου και οι δύο χώρες συνεχίζουν να εξοπλίζονται ασύστολα έτσι θα πάρουν και οι δύο τιμωρίες. Η Ελλάδα θα πάρει μεγαλύτερη τιμωρία (-35) γιατί δαπανεί περισσότερα από ότι η Τουρκία που θα πάρει (-10).

Με τις τιμές που έχουμε επιλέξει πιο πάνω πήραμε ως αποτέλεσμα την πιο κάτω γραφική.

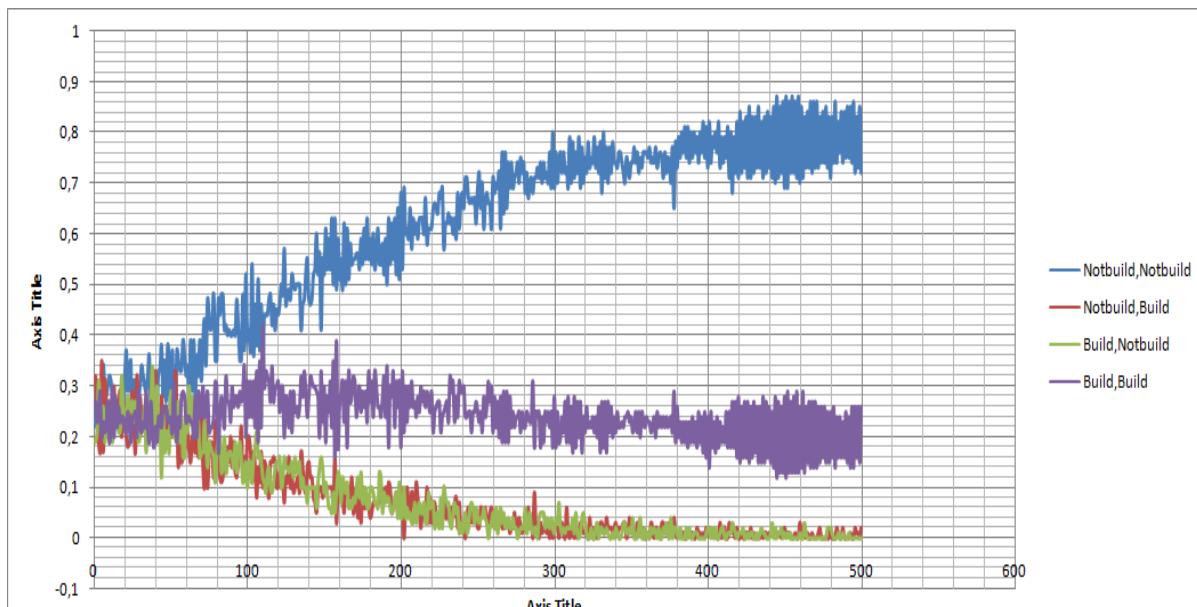
Γραφική Παράσταση 1α: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,9 και οι δύο χώρες εξοπλίζονται ασύστολα. Οι τιμές είναι αντικατοπτρίστηκες του 1974-83.



Ο άξονας των χ είναι οι γύροι στους οποίους εκπαιδεύεται το σύστημα ενώ ο άξονας ψ αντιπροσωπεύει την πιθανότητα εμφάνισης της κάθε κατάστασης σε κάθε γύρο. Οι καταστάσεις είναι αυτές που ορίσαμε και στο ποίο πάνω πίνακα. Ο συμβιβασμός συμβολίζεται με την μπλε γραφική, η Ελλάδα εξοπλίζεται συμβολίζεται με την κόκκινη γραφική, η Τουρκία εξοπλίζεται συμβολίζεται με την πράσινη γραφική ενώ η επίθεση συμβολίζεται με την μωβ γραφική.

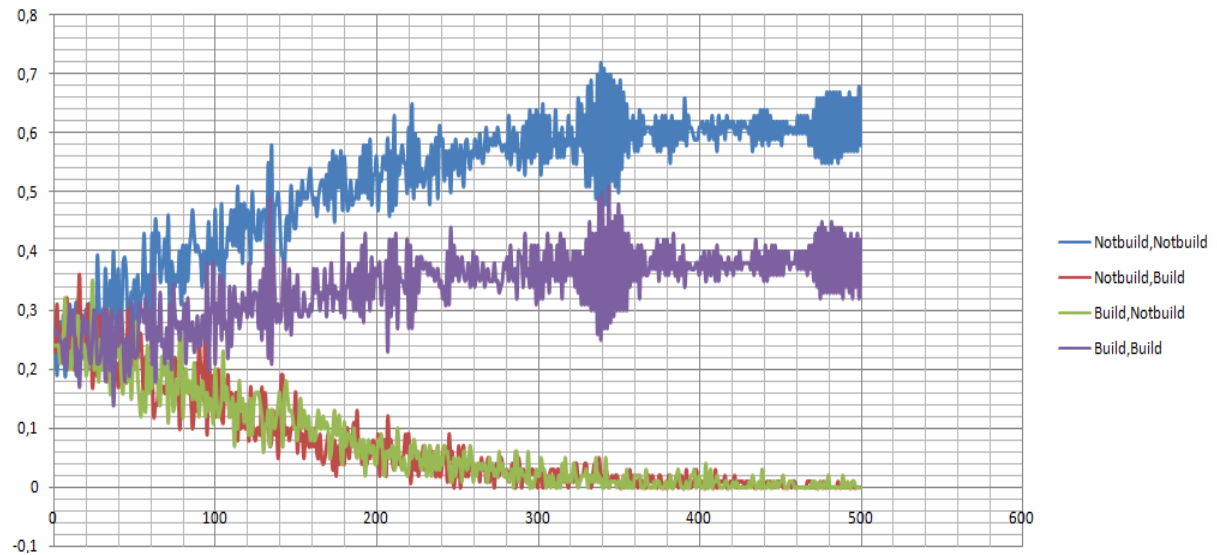
Όπως φαίνεται και από την πιο πάνω γραφική σε περίοδο όπου και οι δύο χώρες κάνουν μεγάλες δαπάνες χρημάτων σε στρατιωτικό εξοπλισμό και η Ελλάδα κάνει ακόμα μεγαλύτερες, υπάρχει μεγαλύτερη δυσκολία να φτάσουν σε αμοιβαίο συμβιβασμό αυτές οι χώρες και αυτό φαίνεται από την γραφική που αντιπροσωπεύει την επίθεση (μωβ), αφού όπως η πιθανότητα εμφάνισης αυξάνεται σταδιακά, άρα ο συμβιβασμός μεταξύ τους δεν είναι και τόσο εύκολος. Αυτό είναι λογικό γιατί εκείνη της εποχή έγινε η εισβολή στην Κύπρο αν η Τουρκία σταματούσε να εξοπλίζεται στρατιωτικά δεν θα είχε καθόλου δύναμη για να κάνει την εισβολή και η Ελλάδα με την σειρά της χωρίς στρατιωτικό εξοπλισμό θα ένιωθε ευάλωτη μπροστά σε μια τόσο μεγάλο πληθυσμό που έχουν οι Τούρκοι. Επομένως σε περιόδους όπως το 74-83 όπου λόγο επιθέσεων ή ακόμα και άμυνας γίνονται τόσο μεγάλες σπατάλες χρημάτων σε στρατιωτική δύναμη είναι αρκετά δύσκολο να πείσεις δύο χώρες να φτάσουν σε αμοιβαίο συμβιβασμό.

Γραφική Παράσταση 1β: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,9 και οι δύο χώρες εξοπλίζονται ασύστολα. Οι τιμές είναι αντικατοπτρίστηκες του 1974-83.



Εδώ έχει γίνει μείωση του ρυθμού μάθησης από 0,9 σε 0,7 όμως οι τιμές των αμοιβών έχουν παραμείνει οι ίδιες. Αυτό που παρατηρούμε είναι ότι δεν υπήρξε μεγάλη διαφορά στην μείωση του ρυθμού μάθησης το σύστημα πάλι μαθαίνει τον συμβιβασμό με λίγη δυσκολία επειδή οι τιμές είναι οι ίδιες και βρισκόμαστε σε περίοδο που και οι δύο χώρες κάνουν μεγάλες δαπάνες σε στρατιωτικό εξοπλισμό. Το μόνο που είναι ελάχιστα διαφορετικό είναι η γραφική του συμβιβασμού (μπλε) που πάει πολύ κοντά στο 0,9 ενώ προηγουμένως πήγαινε κοντά στο 0,8. Καθώς και η γραφική της επίθεσης (μωβ) που φτάνει προς το τέλος των γύρων πολύ κοντά στο 0,3 αντίθετα με την προηγούμενη γραφική της επίθεσης που φτάνει το 0,4. Όμως αυτές οι διαφορές δεν είναι μεγάλες έτσι ο ρυθμός μάθησης είτε μείνει 0,9 είτε μειωθεί σε επιτρεπτά πλαίσια το σύστημα μαθαίνει τον συμβιβασμό στα ίδια επίπεδα στην περίοδο αυτή (μεγάλης δαπάνης και από τις δύο χώρες).

Γραφική Παράσταση 1γ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,7 και οι δύο χώρες εξοπλίζονται ασύστολα. Οι τιμές είναι αντικατοπτρίστηκες του 1974-83.



Εδώ έχει γίνει μείωση του ρυθμού έκπτωσης από 0,9 σε 0,7 όμως οι τιμές των αμοιβών έχουν παραμείνει οι ίδιες. Αυτό που παρατηρούμε είναι ότι όταν μειώνουμε τον ρυθμό έκπτωσης η γραφική του συμβιβασμού δεν φτάνει πολύ ψηλά κυμαίνεται κοντά στο 0,6 και την ακολουθεί αρκετά κοντά η γραφική της επίθεσης που φτάνει μέχρι και 0.5 σε ορισμένους γύρους. Αυτές οι παρατηρήσεις είναι σημαντικές γιατί αποδεικνύεται ότι ο ρυθμός έκπτωσης πρέπει να είναι αρκετά μεγάλος ώστε να πετυχαίνεται ο συμβιβασμός, διαφορετικά θα είναι πολύ δύσκολος. Με ψηλό ρυθμό έκπτωσης το σύστημα μαθαίνει καλύτερα. Επιπλέον σημαντικό είναι να προσθέσουμε ότι ο ρυθμός έκπτωσης πρέπει να είναι μεγάλο για να δίνει ο πράκτορας περισσότερη έμφαση στις μακροπρόθεσμες ανταμοιβές. Καθώς μεγαλώνουμε τον ρυθμό έκπτωσης βλέπουμε ότι οι πράκτορες επιλέγουν το ατομικό συμφέρον που είναι να συνεχίσουν να εξοπλίζονται ασύστολα γιατί αυτό θα τους δώσει μεγάλες άμεσες ανταμοιβές, χάνοντας έτσι την ευκαιρία να πάρουν την μέγιστη συνολική ανταμοιβή μακροπρόθεσμα.

Γραφική Παράσταση 1δ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,7 και οι δύο χώρες εξοπλίζονται ασύστολα. Οι τιμές είναι αντικατοπτρίστικες του 1974-83.



Μειώνοντας και τις δύο παραμέτρους το σύστημα και αφήνοντας και πάλι τις αμοιβές ίδιες, όπως βλέπετε δυσκολεύεται πού να μάθει τον συμβιβασμό αφού οι δύο γραφικές συμβιβασμού και επίθεσης κυμαίνονται στα ίδια επίπεδα 0,6 και 0,5 αντίστοιχα. Άρα πρέπει να έχουμε υψηλές τιμές στις δύο παραμέτρους και ειδικά στο ρυθμό έκπτωσης για πετύχουμε τις δύο χώρες να φτάσουν σε ένα συμβιβασμό και να μην διαλέγουν τις άμεσες μεγάλες ανταμοιβές που δύνονται επιλέγοντας το ατομικό τους συμφέρον αγοράζοντας εξοπλισμό.

4.2 Δεύτερη φάση

Σε αυτή τη φάση θα αναφερθούμε σε χρονικές περιόδους όπου ξοδεύονταν πολύ λίγα σε στρατιωτικό εξοπλισμό. Αυτές οι περίοδοι είναι το 1958-1960 (1961 ήταν περίπου η περίοδος του Τούρκικου πραξικοπήματος), το 1968-1973 (1974 έγινε η εισβολή στη Κύπρο) και το 1989-1997 (όπου ήταν το τέλος του ψυχρού πολέμου), όπου σε αυτές τις περιόδους είχαμε πολύ μικρές δαπάνες και από τις δύο χώρες. Να έχουμε υπόψη μας και πάλι ότι η Ελλάδα σε γενικό βαθμό ξοδεύει περισσότερα σε στρατιωτικό εξοπλισμό από την Τουρκία λόγω του μικρότερου της μεγέθους. Τώρα δεν χρειαζόμαστε μεγάλη διαφορά τιμών μεταξύ των δύο χωρών γιατί και οι δύο χώρες ξοδεύουν περίπου το ίδιο σε χαμηλά επίπεδα τα χρήματα τους για στρατιωτικό εξοπλισμό. Επίσης οι ανταμοιβές και οι τιμωρίες δεν χρειάζεται να είναι τόσο μεγάλες όσο στην προηγούμενη φάση για τον ίδιο λόγο της λιγότερης σπατάλης χρημάτων σε στρατιωτική δύναμη.

Σύμφωνα με αυτά τα γεγονότα ακολουθεί ο πίνακας με τις θέσεις των μερών και μια πιθανή αντιστοίχιση σε καταστάσεις ενός παιγνίου του «δίλλημα του φυλακισμένου». Η Τουρκία παίρνει την αμοιβή με τα κόκκινα γράμματα ενώ η Ελλάδα την αμοιβή με τα μπλε.

Πινάκας 2: Αντιστοίχιση αμοιβών με ρυθμό έκπτωσης 0,9 και ρυθμό μάθησης 0,9 και οι δύο χώρες εξοπλίζονται σε πολύ χαμηλό βαθμό. Οι τιμές αντικατοπτρίζονται στις χρονιές 1958-1960, 1968-1973 και 1989-1997.

Τουρκία	Ελλάδα	
	Συμβιβασμός • Να προβούν και οι δύο χώρες σε μια συμφωνία για να κρατηθεί ειρήνη. • Να μην σπαταλούνται επιπλέον χρήματα σε στρατιωτικό εξοπλισμό. Να βάλουν ένα όριο.	Ελλάδα εξοπλίζεται • Η Ελλάδα εξοπλίζεται σε μικρό βαθμό σύμφωνα με την χρονολογική περίοδο για τον λόγο ότι δεν αισθάνεται απειλή. Μικρό πλήγμα στην οικονομία. • Η Τουρκία δεν εξοπλίζεται περεταίρω. Άρα καθόλου πλήγμα προς την οικονομία της.
	NN2(40,45)	NB2(-10,30)
	Τουρκία εξοπλίζεται • Η Τουρκία εξοπλίζεται σε μικρό βαθμό(λιγότερο από την Ελλάδα) σύμφωνα με την χρονολογική περίοδο για τον λόγο ότι δεν αισθάνεται απειλή. Μικρό πλήγμα στην οικονομία. • Η Τουρκία δεν εξοπλίζεται περεταίρω. Άρα καθόλου πλήγμα προς την οικονομία της.	Επίθεση • Τουρκία και Ελλάδα εξοπλίζονται και οι δύο αλλά σε μικρό, έτσι προκαλούν μικρό πλήγμα στην οικονομία τους(Η Ελλάδα λίγο μεγαλύτερο πλήγμα-μεγαλύτερη δαπάνη). • Ύπαρξη διαμάχης σε μεταγενέστερο στάδιο μεταξύ τους λόγω της υπόνοιας ύπαρξης πολέμου.
	BN2(35,-15)	BB2(-20,-25)

Όπως παρατηρείται από των πίνακα αμοιβών έχουμε ορίσει μια διαφορά +5 μεταξύ των δύο χωρών δηλαδή η Ελλάδα παίρνει +5 μεγαλύτερη αμοιβή ή τιμωρία από την Τουρκία λόγω του ποσοστού 1,23% που ξοδεύει περισσότερα η Ελλάδα. Επίσης η διαφορά είναι μικρότερη από την προηγούμενη περίπτωση επειδή σε αυτή την περίπτωση βρισκόμαστε σε περίοδο όπου ξοδεύουν και οι δύο χώρες λίγα και περίπου τα ίδια με λίγο περισσότερη σπατάλη από την Ελλάδα. Ας μιλήσουμε ξεχωριστά για την κάθε περίπτωση.

Όταν είμαστε στη περίπτωση του συμβιβασμού Η Ελλάδα θα πάρει μια μεγαλύτερη ανταμοιβή (45) από την Τουρκία (40), όμως όπως βλέπεται αυτές οι ανταμοιβές δεν είναι τόσο μεγάλες επειδή βρισκόμαστε σε περίοδο «ησυχίας», και οι δύο χώρες σπαταλούν

λίγα σε στρατιωτικό εξοπλισμό με την Ελλάδα να δαπανεί λίγο περισσότερα. Με αποτέλεσμα το να φτάσουν σε ένα συμβιβασμό η Ελλάδα θα γλιτώσει περισσότερο μέρος των χρημάτων της από την Τουρκία επομένως έχουμε αυτή την διαφορά (+5) στις ανταμοιβές των δύο χωρών.

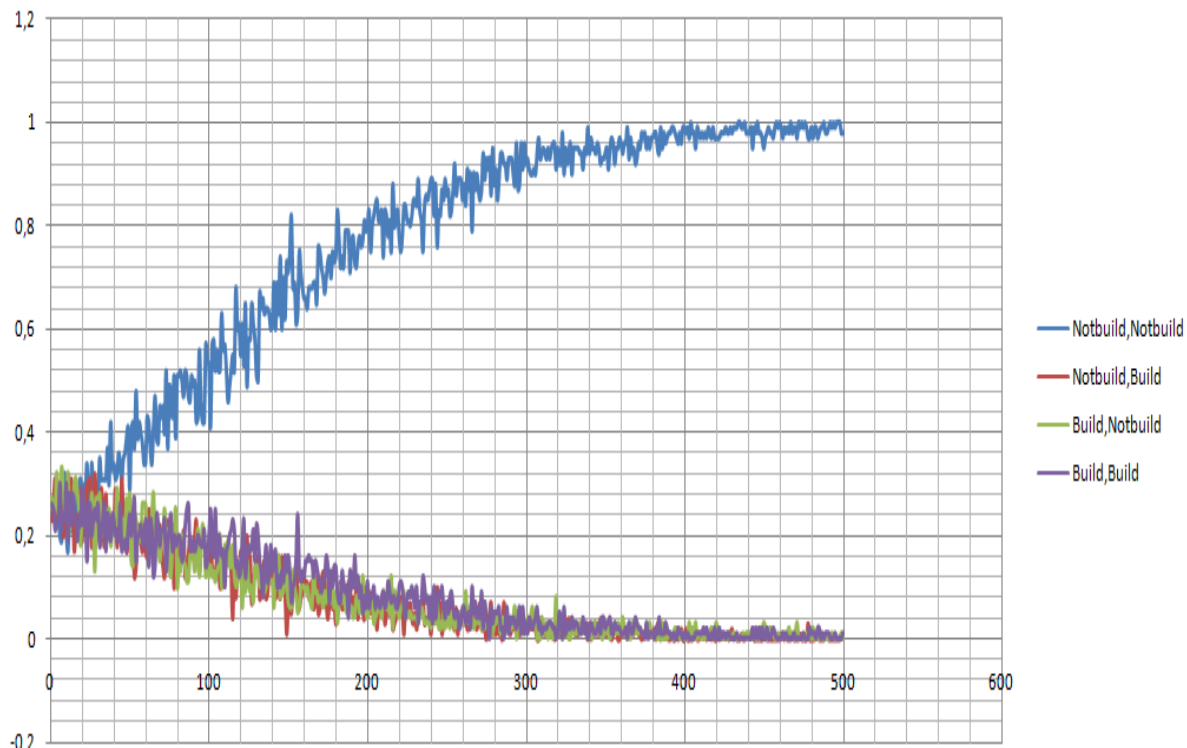
Όταν είμαστε στη περίπτωση που η Ελλάδα ξοδεύει τότε η Ελλάδα θα πάρει αρκετή επιβράβευση 30 γιατί μπορεί να υπερτερεί σε στρατιωτικό εξοπλισμό από την Τουρκία επειδή βρισκόμαστε σε περίοδο ησυχίας (αντίθετα με πριν) όμως πλήττει σε κάποιο βαθμό την οικονομία της γι' αυτό και αυτή η επιβράβευση δεν είναι μεγάλη. Η Τουρκία με την σειρά της παίρνει μια τιμωρία (-10) αφού γίνεται το κορόιδο σε αυτή την περίπτωση της Ελλάδας με τον να μην ξοδέψει πολλά σε εξοπλισμό είναι χαμηλή η τιμωρία επίσης γιατί θυμίζουμε, είμαστε σε περίοδο που ξοδεύονται πολλά λίγα.

Τώρα στην περίπτωση που η Τουρκία εξοπλίζεται θα πάρει μεγαλύτερη ανταμοιβή (35) από εκείνη που πήρε η Ελλάδα όταν εξοπλίστηκε αλλά πάλι είναι σχετικά μεγάλη γιατί βρισκόμαστε σε περίοδο μικρής σπατάλης και δεν έχουν τόσες κακές επιπτώσεις στην οικονομία τους. Επίσης η ανταμοιβή της Τουρκίας είναι +5 σε σχέση με το όταν εξοπλίζεται η Ελλάδα επειδή δεν ξοδεύει σε τόσο βαθμό όσο η Ελλάδα άρα θα έχει μικρότερο πλήγμα η Τουρκία στην οικονομία της άρα θα πάρει μεγαλύτερη σχετικά ανταμοιβή. Τώρα όσο αφορά την Ελλάδα σε αυτή την περίπτωση του εξοπλισμού της Τουρκίας, η Ελλάδα γίνεται κορόιδο και θα πάρει μικρή σχετικά τιμωρία (-15) αφού δεν εξοπλίζεται στρατιωτικά σε μια περίοδο που και οι 2 χώρες εξοπλίζονται σε πολύ μικρό βαθμό.

Τώρα στην τελευταία περίπτωση έχουμε την επίθεση, όπου και οι δύο χώρες συνεχίζουν να εξοπλίζονται ασύστολα έτσι θα πάρουν και οι δύο τιμωρίες. Η Ελλάδα θα πάρει μεγαλύτερη τιμωρία (-25) γιατί δαπανεί περισσότερα από ότι η Τουρκία που θα πάρει (-20).

Με τις τιμές που έχουμε επιλέξει πιο πάνω πήραμε ως αποτέλεσμα την πιο κάτω γραφική.

Γραφική Παράσταση 2α: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,9 και οι δύο χώρες εξοπλίζονται πολύ λίγο. Οι τιμές αντικατοπτρίζονται στις χρονιές 1958-1960, 1968-1973 και 1989-1997.

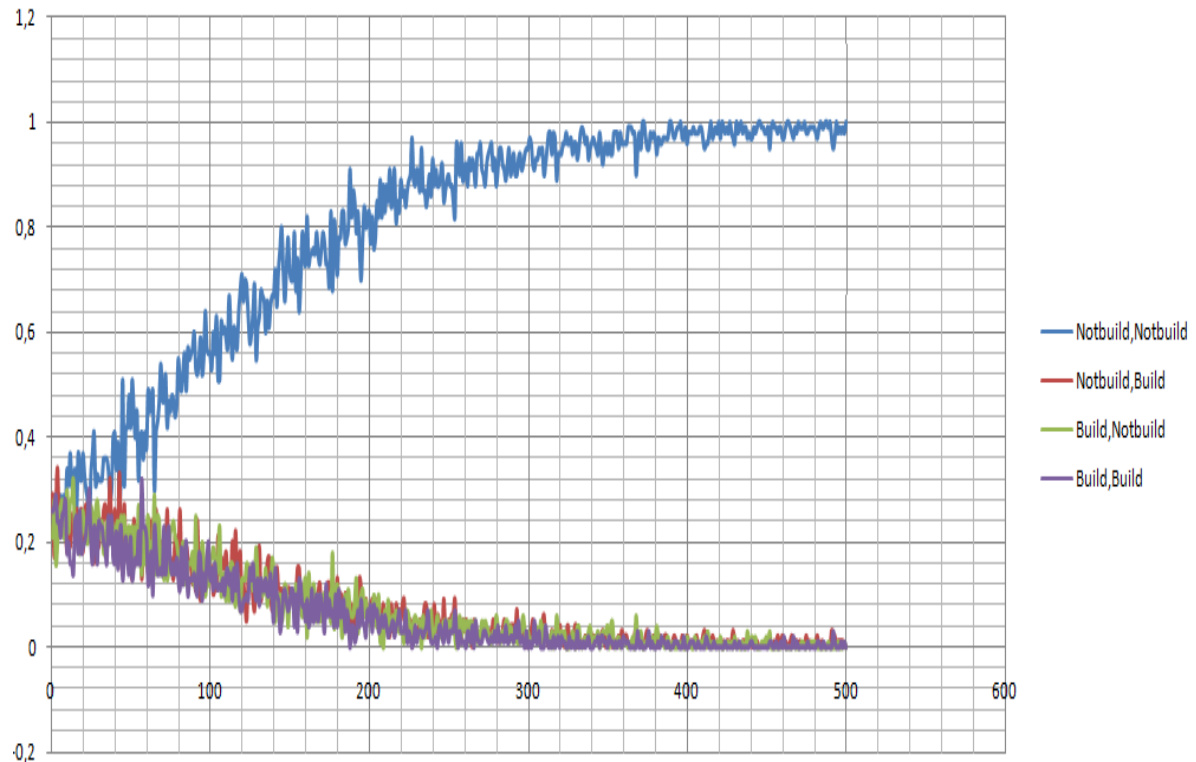


Ο άξονας των χ αντιπροσωπεύει τον αριθμό των γύρων στους οποίους εκπαιδεύεται το σύστημα ενώ ο άξονας ψ αντιπροσωπεύει την πιθανότητα εμφάνισης της κάθε καταστάσεις σε κάθε γύρο. Οι καταστάσεις είναι αυτές που ορίσαμε και στο ποίο πάνω πίνακα. Ο συμβιβασμός συμβολίζεται με την μπλε γραφική, η Ελλάδα εξοπλίζεται συμβολίζεται με την κόκκινη γραφική, η Τουρκία εξοπλίζεται συμβολίζεται με την πράσινη γραφική ενώ η επίθεση συμβολίζεται με την μωβ γραφική.

Βρισκόμαστε σε περίοδο όπου και οι δύο χώρες ξοδεύουν πολύ λίγα σε στρατιωτικό εξοπλισμό. Όπως παρατηρούμε στην πιο πάνω γραφική μπορούν εύκολα οι δύο χώρες να φτάσουν σε αμοιβαίο συμβιβασμό. Αυτό το καταλαβαίνουμε επειδή αρχικά έχουμε τις

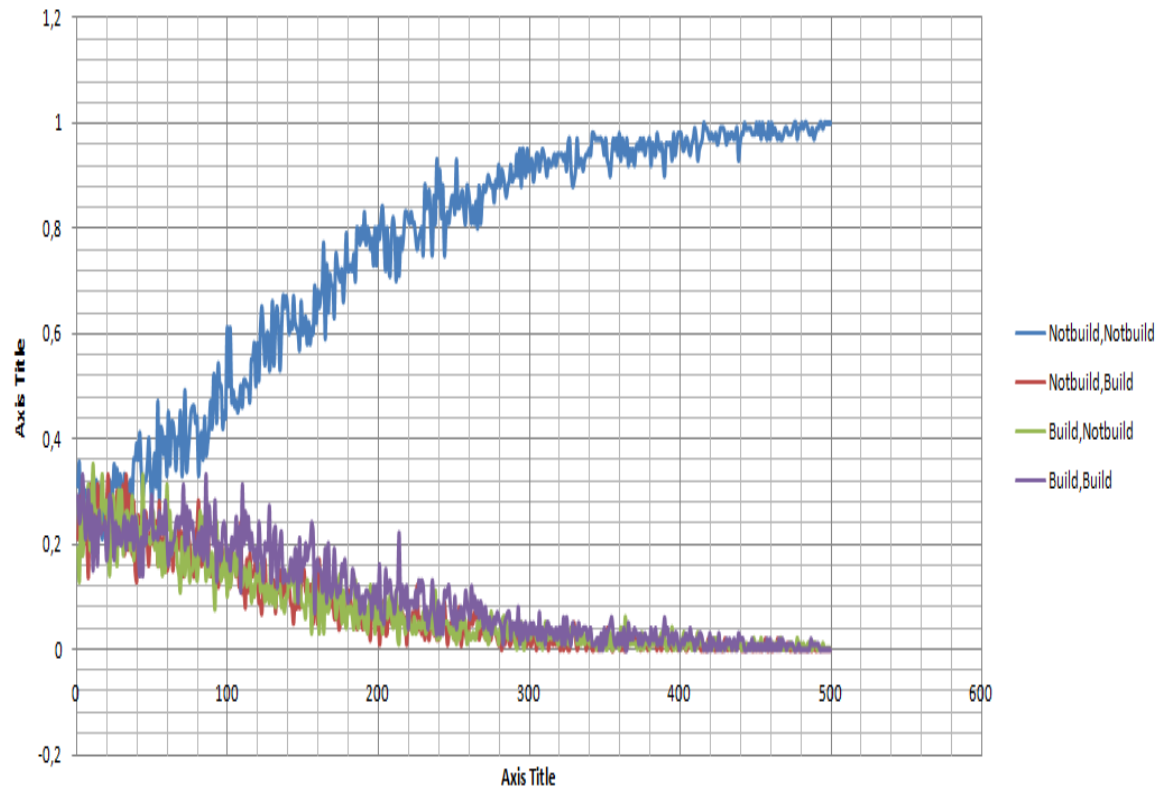
καταστάσεις να βρίσκονται στα ίδια επίπεδα κοντά το 0.2 (από τα 10 που σημαίνει 25% από τα 100) σε πιθανότητα εμφάνισης και σιγά-σιγά βλέπουμε την γραφική του συμβιβασμού (μπλε) να κινείται προς τα πάνω μέχρι την πιθανότητα εμφάνισης 1 ενώ οι υπόλοιπες τρεις γραφικές (που συμβολίζουν τις άλλες τρεις καταστάσεις) κατευθύνονται προς τα κάτω και η πιθανότητα εμφάνισης τους μηδενίζεται. Αυτό σημαίνει ότι σε αυτή την φάση εύκολα οι δύο χώρες μπορούν να φτάσουν σε συμβιβασμό. Αυτό είναι λογικό γιατί στις περιόδους 1958-1960, 1968-1973 και 1989-1997 είχαμε περίοδο ησυχίας, ούτε πόλεμοι υπήρχαν ούτε προετοιμασία για πόλεμο ή για άμυνα έτσι δεν ξοδεύονταν μεγάλοι όγκοι χρημάτων για στρατιωτικοί δυνάμει, έτσι είναι πιο εύκολο να πείσεις δύο χώρες που δεν βρίσκονται σε καμία διαμάχη να φτάσουν σε συμβιβασμό.

Γραφική Παράσταση 2β: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,9 και οι δύο χώρες εξοπλίζονται πολύ λίγο. Οι τιμές αντικατοπτρίζονται στις χρονιές 1958-1960, 1968-1973 και 1989-1997.



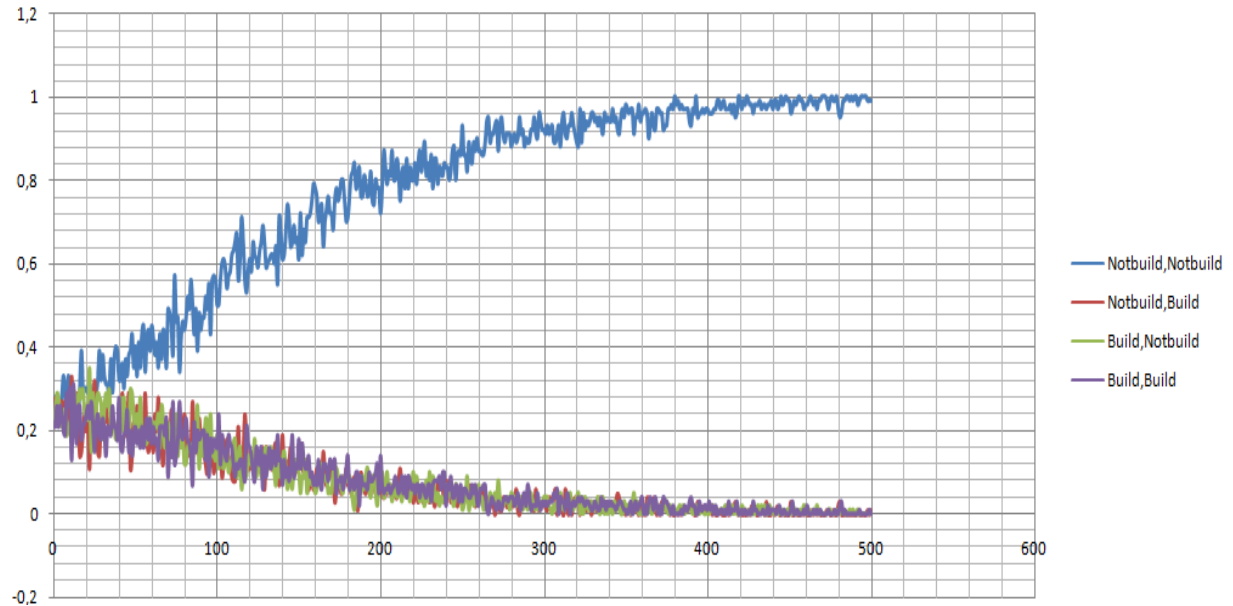
Μειώνοντας το ρυθμό μάθησης δεν βλέπουμε κάποια αισθητή διαφορά στο σύστημα. Μαθαίνει να φτάνει γρήγορα σε συμβιβασμό σε περίοδο που και οι δύο χώρες ξοδεύουν ελάχιστα σε στρατιωτικό εξοπλισμό.

Γραφική Παράσταση 2γ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,7 και οι δύο χώρες εξοπλίζονται πολύ λίγο. Οι τιμές αντικατοπτρίζονται στις χρονιές 1958-1960, 1968-1973 και 1989-1997.



Παρατηρούμε ότι με την μείωση του ρυθμού έκπτωσης η γραφική της επίθεσης (μωβ) παίρνει πιο μεγάλες τιμές περίπου ως τα μισά των γύρων ενώ μετά καταλήγει και πάλι στο 0. Όμως όσο μειώνουμε το ρυθμό έκπτωσης τότε ο πράκτορας θα γίνεται μυωπικός και θα τον ενδιαφέρουν μόνο οι άμεσες ανταμοιβές κάτι που εμείς δεν θέλουμε. Εμείς θέλουμε να δώσουμε περισσότερη αξία στη μακροπρόθεσμες ανταμοιβές. Μειώνοντας το ρυθμό έκπτωσης δεν θα το πετυχαίναμε αυτό. Συνεπώς θέλουμε μεγάλο ρυθμό έκπτωσης.

Γραφική Παράσταση 2δ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,7 και οι δύο χώρες εξοπλίζονται πολύ λίγο. Οι τιμές αντικατοπτρίζονται στις χρονιές 1958-1960, 1968-1973 και 1989-1997.



Η μείωση των παραμέτρων όπως φαίνεται δεν επηρεάζει τα αποτελέσματα έχουμε τα ίδια αποτελέσματα που είχαμε και με τις αρχικές τιμές των παραμέτρων. Όμως όπως είπαμε και πριν ο ρυθμός έκπτωσης πρέπει να είναι αρκετά ψηλός για να βλέπουν οι παίχτες μακροπρόθεσμα.

4.3 Τρίτη φάση

Τώρα θα αναφερθούμε σε μια χρονολογία όπου η Ελλάδα σπαταλούσε μεγαλύτερο ποσοστό χρημάτων σε στρατιωτικό εξοπλισμό σε σχέση με την Τουρκία. Αναφέρομαι στην χρονιά του 1984-1988 (μέχρι το 1983 η Τουρκία είχε στρατιωτικό καθεστώς πριν αυτό αλλάξει). Να έχουμε υπόψη μας και πάλι ότι η Ελλάδα σε γενικό βαθμό ξοδεύει περισσότερα σε στρατιωτικό εξοπλισμό από την Τουρκία λόγω του μικρότερου της μεγέθους.

Τώρα χρειαζόμαστε μεγαλύτερη διαφορά τιμών μεταξύ των δύο χωρών γιατί έχουμε την Ελλάδα να εξοπλίζεται σε αυτή την περίοδο ενώ η Τουρκία εξοπλίζεται σε πολύ χαμηλά επίπεδα και σε γενικό βαθμό ξέρουμε ότι η Ελλάδα εξοπλίζεται κατά 1,23% περισσότερο από την Τουρκία. Έτσι συμπερασματικά η Ελλάδα ξοδεύει πολύ περισσότερα σε αυτή την περίοδο έναντι της Τουρκίας που ξοδεύει πολύ λίγα έτσι θα ορίσουμε μια διαφορά τιμών +30 για να δήξει αυτή την διαφορά που θέλουμε. Επίσης οι ανταμοιβές και οι τιμωρίες θα είναι αρκετά μεγάλες για την Ελλάδα η οποία εξοπλίζεται σε μεγαλύτερους ρυθμούς αυτή την περίοδο.

Σύμφωνα με αυτά τα γεγονότα ακολουθεί ο πίνακας με τις θέσεις των μερών και μια πιθανή αντιστοίχιση σε καταστάσεις ενός παιγνίου του «δύλλημα του φυλακισμένου». Η Τουρκία παίρνει την αμοιβή με τα κόκκινα γράμματα ενώ η Ελλάδα την αμοιβή με τα μπλε.

Πίνακας 3: Αντιστοίχιση αμοιβών με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,9 η Ελλάδα εξοπλίζεται σε πολύ μεγαλύτερο επίπεδο από την Τουρκία. Οι τιμές αντικατοπτρίζονται στην χρονιά 1984-1988.

Τουρκία	Ελλάδα	
	Συμβιβασμός • Να βρεθεί ένα ισοζύγιο ώστε η μια χώρα να μην ξοδεύει περισσότερο σε στρατιωτικό εξοπλισμό από την άλλη. • Να υπάρξει συμφωνία που να τους συμμορφώνει για να μην υπάρξει πλήγμα στην οικονομία των 2 χωρών.	Ελλάδα εξοπλίζεται • Η Ελλάδα εξοπλίζεται σε μεγάλο βαθμό σύμφωνα με την χρονολογική περίοδο Τεράστιο πλήγμα στην οικονομία. • Η Τουρκία δεν εξοπλίζεται σχεδόν καθόλου με αποτέλεσμα να μην υποστεί κανένα πλήγμα στην οικονομία της.
	NN3(30,60)	NB3(-30,10)
	Τουρκία εξοπλίζεται • Η Τουρκία εξοπλίζεται σε μικρό βαθμό(λιγότερο από την Ελλάδα) σύμφωνα με την χρονολογική περίοδο για τον λόγο ότι δεν αισθάνεται απειλή. Μικρό πλήγμα στην οικονομία. • Η Τουρκία δεν εξοπλίζεται περεταίρω. Άρα καθόλου πλήγμα προς την οικονομία της.	Επίθεση • Τουρκία και Ελλάδα εξοπλίζονται και οι δύο αλλά σε μικρό επίπεδο, έτσι προκαλούν μικρό πλήγμα στην οικονομία τους (Η Ελλάδα λίγο μεγαλύτερο πλήγμα-μεγαλύτερη δαπάνη). • Ύπαρξη διαμάχης σε μεταγενέστερο στάδιο μεταξύ τους λόγω της υπόνοιας ύπαρξης πολέμου.
	BN3(40,-60)	BB3(0,-30)

Όπως αναφέραμε και πιο πριν ορίσαμε μια διαφορά τιμών +30 μεταξύ των δύο χωρών δηλαδή η Ελλάδα παίρνει +30 μεγαλύτερη ανταμοιβή ή τιμωρία από την Τουρκία λόγω της πολύ μεγαλύτερης δαπάνης που κάνει η Ελλάδα αυτή την περίοδο με το να εξοπλίζεται σε πολύ μεγαλύτερους ρυθμούς από κάθε άλλη φορά σε αντίθεση με την Τουρκία που εξοπλίζεται σε ελάχιστα επίπεδα. Ας μιλήσουμε ξεχωριστά για την κάθε περίπτωση.

Όταν είμαστε στη περίπτωση του συμβιβασμού η Ελλάδα θα πάρει μια μεγαλύτερη μια μεγάλη ανταμοιβή (60) σε σχέση με την Τουρκία (30). Αυτό συμβαίνει επειδή αν η Ελλάδα

σταματήσει τις μεγάλες δαπάνες σε στρατιωτικό εξοπλισμό θα γλιτώσει πολλά χρήματα αυτή την περίοδο (που ξοδεύει υπερβολικά) και αυτό θα έχει θετικές επιπτώσεις στην οικονομία της. Η Τουρκία με το συμβιβασμό σε αυτή την περίοδο στην οποία δεν εξοπλιζόταν πολύ στρατιωτικά θα πάρει μικρότερη ανταμοιβή από την Ελλάδα, γιατί δεν έχει να γλιτώσει μεγάλους όγκους χρημάτων λόγω των λίγο που ξόδευε. Γι' αυτό το λόγο έχουμε ορίσει και την διαφορά τιμών +30.

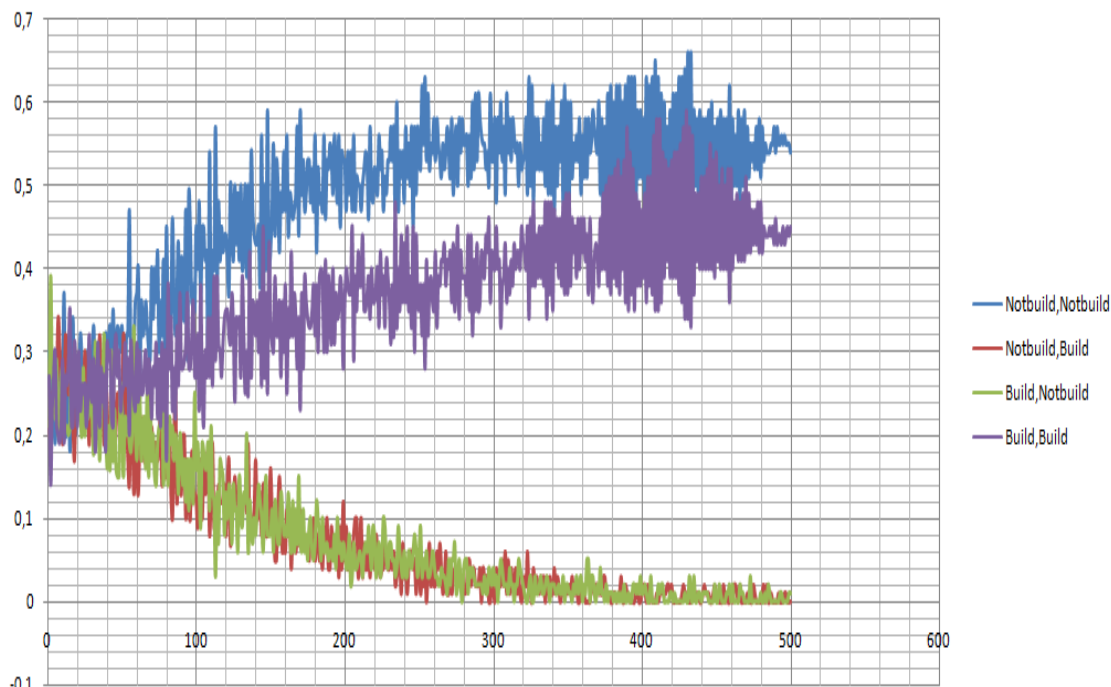
Όταν είμαστε στη περίπτωση που η Ελλάδα εξοπλίζεται τότε η Ελλάδα θα πολύ μικρή επιβράβευση (10) παρόλο που υπερσχύει η στρατιωτική της δύναμη θα προκαλέσει σοβαρές επιπτώσεις στην οικονομία της σε αυτή την περίοδο, για αυτό το λόγο παίρνει και τόσο χαμηλή επιβράβευση. Η Τουρκία με την σειρά της παίρνει μια τιμωρία (-30) αφού γίνεται το κορόιδο σε αυτή την περίπτωση της Ελλάδας με τον να μην ξοδέψει πολλά σε εξοπλισμό, η τιμωρία της σχετικά χαμηλή η γιατί ξοδεύει πολύ λίγα σε στρατιωτικό εξοπλισμό έτσι η οικονομία της δεν έχει πτώση.

Τώρα στην περίπτωση που η Τουρκία εξοπλίζεται θα πάρει μεγαλύτερη ανταμοιβή (40) από εκείνη που πήρε η Ελλάδα όταν εξοπλίστηκε. Η ανταμοιβή της Τουρκίας είναι +30 σε σχέση με το όταν εξοπλίζεται η Ελλάδα επειδή δεν ξοδεύει σε τόσο βαθμό όσο η Ελλάδα άρα θα έχει μικρότερο πλήγμα η Τουρκία στην οικονομία της άρα θα πάρει μεγαλύτερη σχετικά ανταμοιβή. Τώρα όσο αφορά την Ελλάδα σε αυτή την περίπτωση του εξοπλισμού της Τουρκίας, γίνεται κορόιδο και θα πάρει πιο μεγάλη τιμωρία (-60) αφού δεν εξοπλίζεται στρατιωτικά σε μια περίοδο όπου θα έπρεπε να εξοπλιζόταν για να υπερσχύσει σε στρατιωτική δύναμη.

Τώρα στην τελευταία περίπτωση έχουμε την επίθεση, όπου και οι δύο χώρες συνεχίζουν να εξοπλίζονται με γοργούς ρυθμούς έτσι θα πάρουν και οι δύο χώρες τιμωρία. Η Ελλάδα θα πάρει μεγαλύτερη τιμωρία (-30) γιατί δαπανεί πολύ περισσότερα σε αυτή την περίοδο από ότι η Τουρκία που θα πάρει (0).

Με τις τιμές που έχουμε επιλέξει πιο πάνω πήραμε ως αποτέλεσμα την πιο κάτω γραφική.

Γραφική Παράσταση 3α: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,9 η Ελλάδα εξοπλίζεται σε πολύ μεγαλύτερο επίπεδο από την Τουρκία. Οι τιμές αντικατοπτρίζονται στην χρονιά 1984-1988.

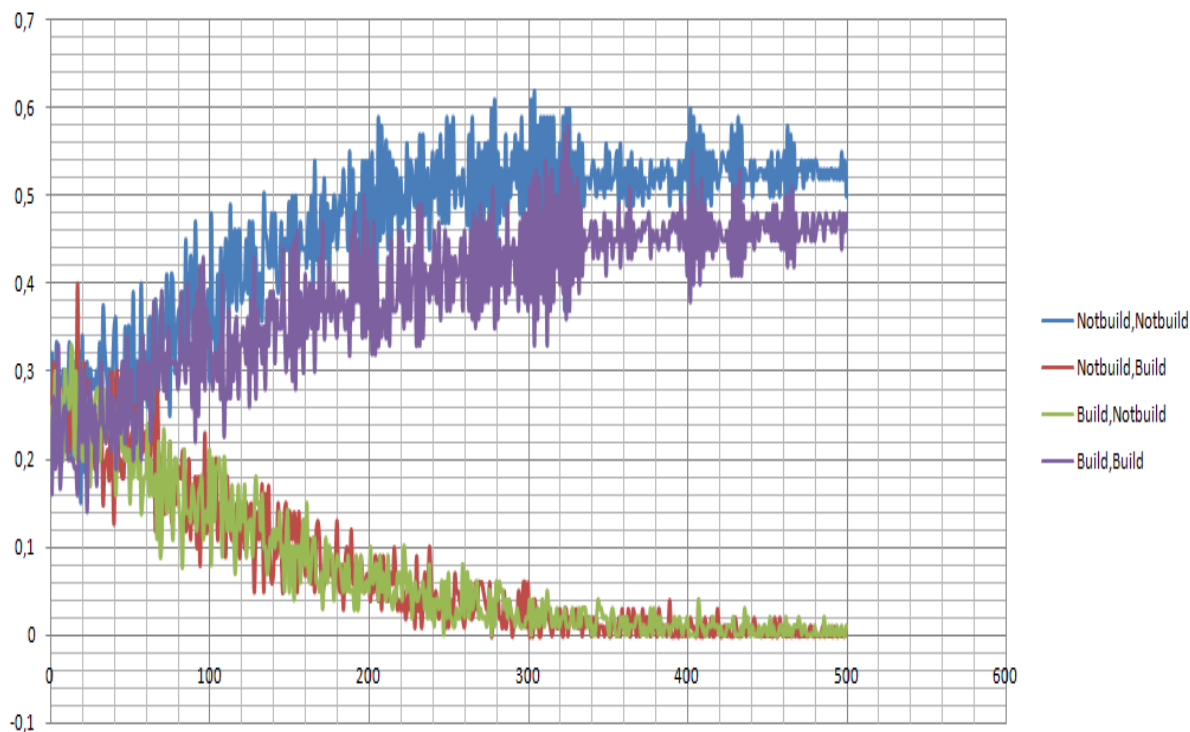


Ο χ άξονας αντιπροσωπεύει γύρους στους οποίους εκπαιδεύεται το σύστημα ενώ ο ψ άξονας αντιπροσωπεύει την πιθανότητα εμφάνισης της κάθε καταστάσεις σε κάθε γύρο. Οι καταστάσεις είναι αυτές που ορίσαμε και στο ποίο πάνω πίνακα. Ο συμβιβασμός συμβολίζεται με την μπλε γραφική, η Ελλάδα εξοπλίζεται συμβολίζεται με την κόκκινη γραφική, η Τουρκία εξοπλίζεται συμβολίζεται με την πράσινη γραφική ενώ η επίθεση συμβολίζεται με την μωβ γραφική.

Όπως είπαμε πιο πάνω βρισκόμαστε σε περίοδο όπου και η Ελλάδα ξοδεύει πολύ περισσότερα σε στρατιωτικό εξοπλισμό ενώ η Τουρκία ξοδεύει σε πολύ χαμηλά επίπεδα. Όπως παρατηρούμε στην πιο πάνω γραφική οι δύο χώρες δεν μπορούν να φτάσουν εύκολα σε ένα συμβιβασμό. Αυτό το καταλαβαίνουμε επειδή αρχικά έχουμε τις καταστάσεις να

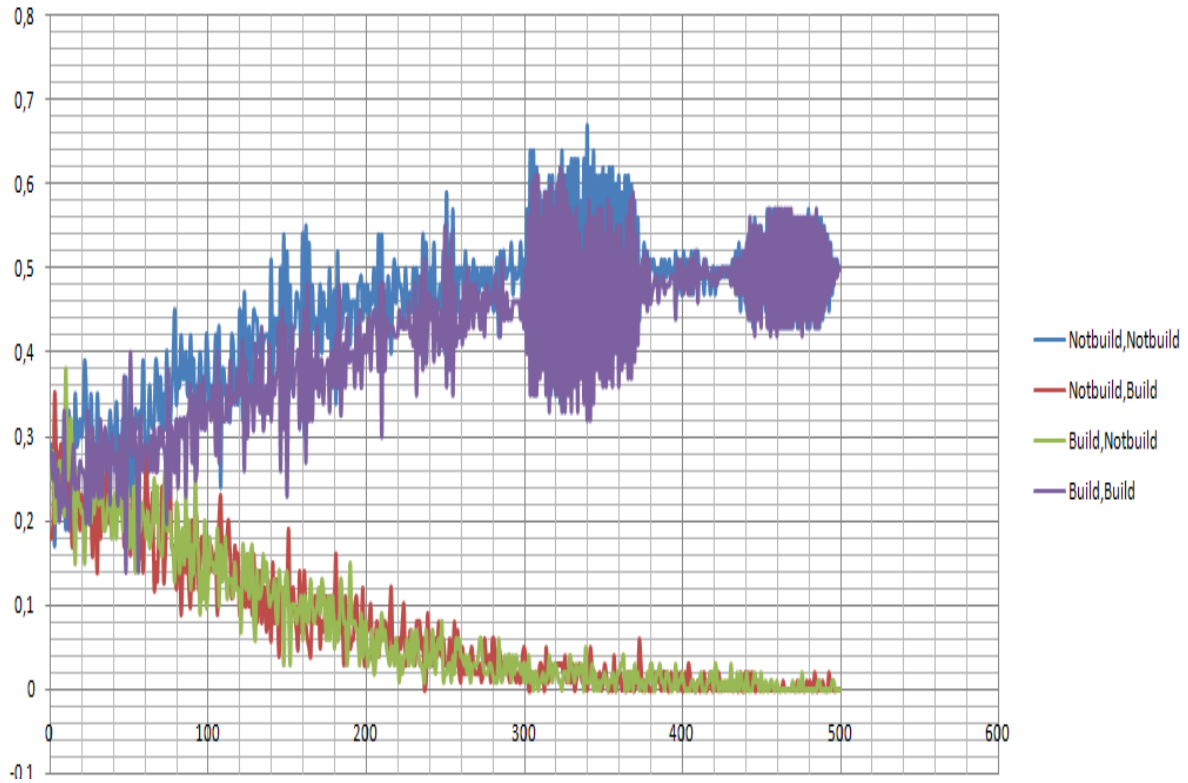
βρίσκονται στα ίδια επίπεδα κοντά το 0.2 (από τα 10 που σημαίνει 25% από τα 100) σε πιθανότητα εμφάνισης και σιγά-σιγά βλέπουμε την γραφική του συμβιβασμού (μπλε) να κινείται προς τα πάνω μέχρι την πιθανότητα εμφάνισης 1, έπειτα βλέπουμε την γραφική της επίθεσης να πλησιάζει επικίνδυνα προς τη πιθανότητα εμφάνισης 1 ενώ οι υπόλοιπες δύο γραφικές (που συμβολίζουν τις άλλες δύο καταστάσεις) κατευθύνονται προς τα κάτω και η πιθανότητα εμφάνισης τους μηδενίζεται. Αυτό σημαίνει ότι σε αυτή την φάση δύσκολα οι δύο χώρες μπορούν να φτάσουν σε συμβιβασμό με την κατάσταση της επίθεσης να παίρνει πολύ ψηλές πιθανότητες εμφάνισης κατά την διάρκεια των γύρων. Αυτό είναι λογικό γιατί στην περίοδο 1984-1988 είχαμε την Ελλάδα να εξοπλίζεται ασύστολα, επομένως και σε αυτή την περίπτωση είναι δύσκολος ο συμβιβασμός όμως επιτυχαίνεται αφού η γραφική που τον συμβολίζει παίρνει πιο ψηλές τιμές από την γραφική της επίθεσης.

Γραφική Παράσταση 3β: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,9 η Ελλάδα εξοπλίζεται σε πολύ μεγαλύτερο επίπεδο από την Τουρκία. Οι τιμές αντικατοπτρίζονται στην χρονιά 1984-1988.



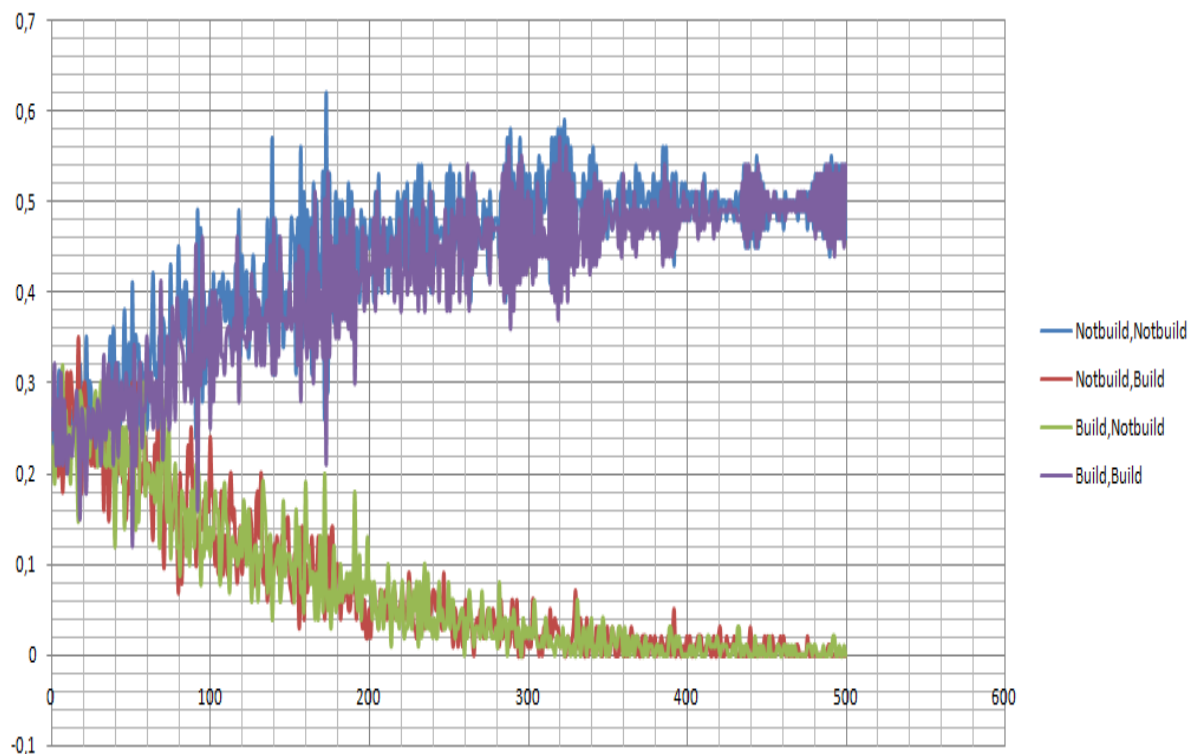
Μειώνοντας το ρυθμό μάθησης (τιμές των ανταμοιβών παραμένουν οι ίδιες) δεν έχουμε κάποια αισθητή διαφορά σε αυτή την περίοδο επειδή η Ελλάδα ξεοδεύει πολύ περισσότερα από την Τουρκία ο συμβιβασμός καθιστάτε πολύ δύσκολος αν και η γραφική του συμβιβασμού (μπλε) παίρνει λίγο πιο ψηλές τιμές (λίγο μεγαλύτερη πιθανότητα εμφάνισης) από την γραφική της επίθεσης (μωβ). Επομένως η μείωση του ρυθμού μάθησης δεν επηρεάζει την μάθηση σε μεγάλο βαθμό.

Γραφική Παράσταση 3γ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,7 η Ελλάδα εξοπλίζεται σε πολύ μεγαλύτερο επίπεδο από την Τουρκία. Οι τιμές αντικατοπτρίζονται στην χρονιά 1984-1988.



Όπως παρατηρούμε από την γραφική παράσταση, μειώνοντας τον ρυθμό έκπτωσης έχοντας τις ίδιες τιμές στο πίνακα τιμών τα αποτελέσματα που παίρνουμε είναι χειρότερα από πριν. Η εμφάνιση της κατάστασης του συμβιβασμού και της επίθεσης βρίσκονται στα ίδια επίπεδα όπως φαίνεται από τις δύο γραφικές. Συνεπώς αυτό συμβαίνει γιατί μειώνοντας τον ρυθμό έκπτωσης η παίχτες ενδιαφέρονται μόνο για τις βραχυπρόθεσμες ανταμοιβές. Φαίνεται ότι επιλέγουν το ατομικό συμφέρον που θα τους επιφέρει μεγάλες άμεσες ανταμοιβές επιλέγοντας να συνεχίζουν να εξοπλίζονται ασύστολα. Γι' αυτό καλά θα ήταν να δώσουμε μεγάλη τιμή στο ρυθμό έκπτωσης ώστε η παίχτες να ενδιαφέρονται για την μέγιστη συνολική μακροπρόθεσμη ανταμοιβή που θα επέλθει αν δουν το κοινό καλό και επιλέξουν να συμβιβαστούν.

Γραφική Παράσταση 3δ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,7 η Ελλάδα εξοπλίζεται σε πολύ μεγαλύτερο επίπεδο από την Τουρκία. Οι τιμές αντικατοπτρίζονται στην χρονιά 1984-1988.



Όπως παρατηρούμε με την μείωση και των δύο παραμέτρων παίρνουμε χειρότερα αποτελέσματα αφού πάλι ο συμβιβασμός και η επίθεση έχουν τις ίδιες πιθανότητες εμφάνισης. Επομένως χρειαζόμαστε υψηλό ρυθμό έκπτωσης για να βλέπουν οι παίχτες μακροπρόθεσμα και να επιζητούν την μέγιστη μακροπρόθεσμη ανταμοιβή που δίνεται από την επιλογή του συμβιβασμού αφού είναι η καλύτερη επιλογή και για τους δύο. Καθώς και υψηλό ρυθμό μάθησης για να μαθαίνουν γρήγορα.

4.4 Τέταρτη φάση

Σύμφωνα με την χρονολογία 1961-1967 η Τουρκία εξοπλίζεται με πιο μεγάλο ρυθμό από ότι η Ελλάδα, μια περίοδος όπου γίνεται το στρατιωτικό πραξικόπημα από την Τουρκία. Να έχουμε υπόψη μας και πάλι ότι η Ελλάδα σε γενικό βαθμό ξοδεύει περισσότερο σε στρατιωτικό εξοπλισμό από την Τουρκία λόγω του μικρότερου της μεγέθους, όμως τώρα θα λάβουμε υπόψη και ότι η Τουρκία σε αυτή τη περίοδο ξοδεύει περισσότερο. Άρα όσο αφορά την διαφορά μεταξύ των τιμών Ελλάδας Τουρκίας θα είναι +8. Δεν θα είναι μεγάλη γιατί παρόλο που η Τουρκία εξοπλίζεται περισσότερο αυτή την φάση η Ελλάδα σε γενικό βαθμό ξοδεύει πολλά σε στρατιωτικό εξοπλισμό λόγω του μεγέθους της, έτσι η μεγάλη σπατάλη της Τουρκίας αντισταθμίζεται κάπως από την γενική μεγάλη σπατάλη της Ελλάδα γι' αυτό και θέσαμε τόσο μικρή διαφορά μεταξύ των αμοιβών.

Σύμφωνα με αυτά τα γεγονότα ακολουθεί ο πίνακας με τις θέσεις των μερών και μια πιθανή αντιστοίχιση σε καταστάσεις ενός παιγνίου του «δίλλημα του φυλακισμένου». Η Τουρκία παίρνει την αμοιβή με τα κόκκινα γράμματα ενώ η Ελλάδα την αμοιβή με τα μπλε.

Πίνακας 4: Αντιστοίχιση αμοιβών με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,9 η Τουρκία εξοπλίζεται σε λίγο περισσότερο επίπεδο από ότι η Ελλάδα. Οι τιμές αντικατοπτρίζονται στη χρονιά 1961-1967.

Τουρκία	Ελλάδα	
	Συμβιβασμός	Ελλάδα εξοπλίζεται
	• Να βρεθεί ένα ισοζύγιο ώστε η μια χώρα να μην ξοδεύει περισσότερο σε στρατιωτικό εξοπλισμό από την άλλη. • Να υπάρξει συμφωνία που να τους συμμορφώνει για να μην υπάρξει πλήγμα στην οικονομία των 2 χωρών.	• Η Ελλάδα εξοπλίζεται σε μεγάλο βαθμό σύμφωνα με την χρονολογική περίοδο Τεράστιο πλήγμα στην οικονομία. • Η Τουρκία δεν εξοπλίζεται σχεδόν καθόλου με αποτέλεσμα να μην υποστεί κανένα πλήγμα στην οικονομία της.
	NN4(55,48)	NB4(-40,23)
Τουρκία	Τουρκία εξοπλίζεται	Επίθεση
	• Η Τουρκία εξοπλίζεται σε μικρό βαθμό(λιγότερο από την Ελλάδα) σύμφωνα με την χρονολογική περίοδο για τον λόγο ότι δεν αισθάνεται απειλή. Μικρό πλήγμα στην οικονομία. • Η Τουρκία δεν εξοπλίζεται περεταίρω. Άρα καθόλου πλήγμα προς την οικονομία της.	• Τουρκία και Ελλάδα εξοπλίζονται και οι δύο αλλά σε μικρό επίπεδο, έτσι προκαλούν μικρό πλήγμα στην οικονομία τους (Η Ελλάδα λίγο μεγαλύτερο πλήγμα-μεγαλύτερη δαπάνη). • Ύπαρξη διαμάχης σε μεταγενέστερο στάδιο μεταξύ τους λόγω της υπόνοιας ύπαρξης πολέμου.
	BN4(15,-32)	BB4(-10,-3)

Όπως αναφέραμε και προηγουμένως ορίσαμε μια πολύ μικρή διαφορά τιμών +8 μεταξύ των δύο χωρών δηλαδή η Τουρκία παίρνει +8 μεγαλύτερη ανταμοιβή ή τιμωρία από την Τουρκία. Αυτό συμβαίνει επειδή παρόλο που βρισκόμαστε σε περίοδο που η Τουρκία εξοπλίζεται περισσότερο από την Ελλάδα, η Ελλάδα με την σειρά τις πάντα εξοπλίζεται 1,23% περισσότερο από την Τουρκία έτσι η μεγάλες δαπάνες που κάνει η Τουρκία αντισταθμίζονται κατά ένα μέρος από τις γενικές δαπάνες που κάνει πάντα η Ελλάδα, όμως η Τουρκία σε τελικό βαθμό ξοδεύει λίγα περισσότερα λαμβάνοντας υπόψη αυτά που είπαμε. Ας μιλήσουμε ξεχωριστά για την κάθε περίπτωση.

Όταν είμαστε στη περίπτωση του συμβιβασμού η Τουρκία θα πάρει μια μεγαλύτερη ανταμοιβή (55) σε σχέση με την Τουρκία (48). Η διαφορά είναι μικρή για τους λόγους που

αναφέραμε πιο πάνω και αυτό έχει ως αποτέλεσμα η Τουρκία να γλιτώνει λίγα περισσότερα λεφτά (από ότι η Ελλάδα) από τις δαπάνες που έκανε αν δεχόταν το συμβιβασμό. Με αποτέλεσμα να έχουν και οι δύο χώρες θετικές επιπτώσεις στην οικονομία τους με την Τουρκία να έχει λίγο καλύτερες.

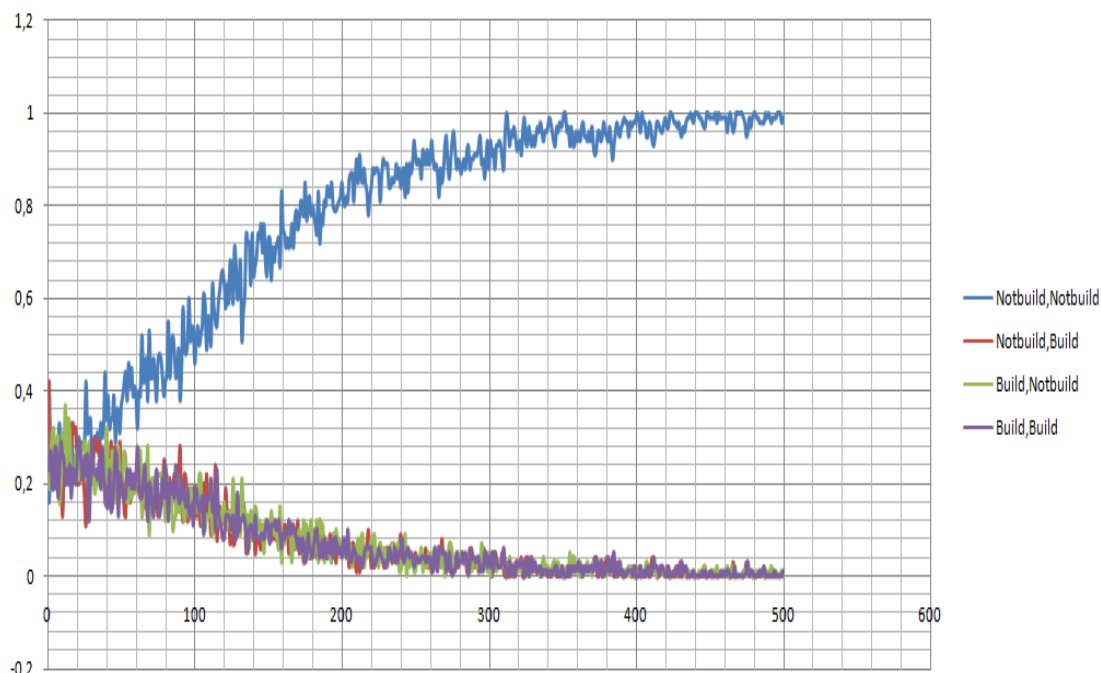
Όταν είμαστε στη περίπτωση που η Ελλάδα εξοπλίζεται τότε η Ελλάδα θα πάρει λίγη επιβράβευση (23) παρόλο που υπερισχύει η στρατιωτική της δύναμη θα προκαλέσει σοβαρές επιπτώσεις στην οικονομία της σε αυτή την περίοδο. Η Τουρκία με την σειρά της παίρνει μια τιμωρία (-40) αφού γίνεται το κορόιδο σε αυτή την περίπτωση της Ελλάδας με τον να μην ξοδεύει πολλά σε εξοπλισμό.

Τώρα στην περίπτωση που η Τουρκία εξοπλίζεται θα πάρει μικρότερη ανταμοιβή (15) από εκείνη που πήρε η Ελλάδα όταν εξοπλίστηκε. Η ανταμοιβή της Ελλάδας είναι +8 σε σχέση με το όταν εξοπλίζεται η Τουρκία επειδή δεν ξοδεύει σε τόσο βαθμό όσο η Ελλάδα σύμφωνα με αυτά που είπαμε πιο πάνω, άρα θα έχει μικρότερο πλήγμα η Ελλάδα στην οικονομία της άρα θα πάρει μεγαλύτερη σχετικά ανταμοιβή. Τώρα όσο αφορά την Ελλάδα σε αυτή την περίπτωση του εξοπλισμού της Τουρκίας, γίνεται κορόιδο και θα πάρει πιο μικρή τιμωρία (-32) από ότι η τιμωρία που παίρνει η Τουρκία όταν εξοπλίζεται η Ελλάδα αφού δεν εξοπλίζεται στρατιωτικά σε μια περίοδο όπου θα έπρεπε να εξοπλιζόταν για να υπερисχύσει σε στρατιωτική δύναμη.

Τώρα στην τελευταία περίπτωση έχουμε την επίθεση, όπου και οι δύο χώρες συνεχίζουν να εξοπλίζονται με γοργούς ρυθμούς έτσι θα πάρουν και οι δύο χώρες τιμωρία. Η Τουρκία θα πάρει μεγαλύτερη τιμωρία (-10) γιατί δαπανεί λίγο περισσότερα σε αυτή την περίοδο από ότι η Τουρκία που θα πάρει (-3).

Με τις τιμές που έχουμε επιλέξει πιο πάνω πήραμε ως αποτέλεσμα την πιο κάτω γραφική.

Γραφική Παράσταση 4α: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,9 η Τουρκία εξοπλίζεται σε λίγο περισσότερο επίπεδο από ότι η Ελλάδα. Οι τιμές αντικατοπτρίζονται στη χρονιά 1961-1967.

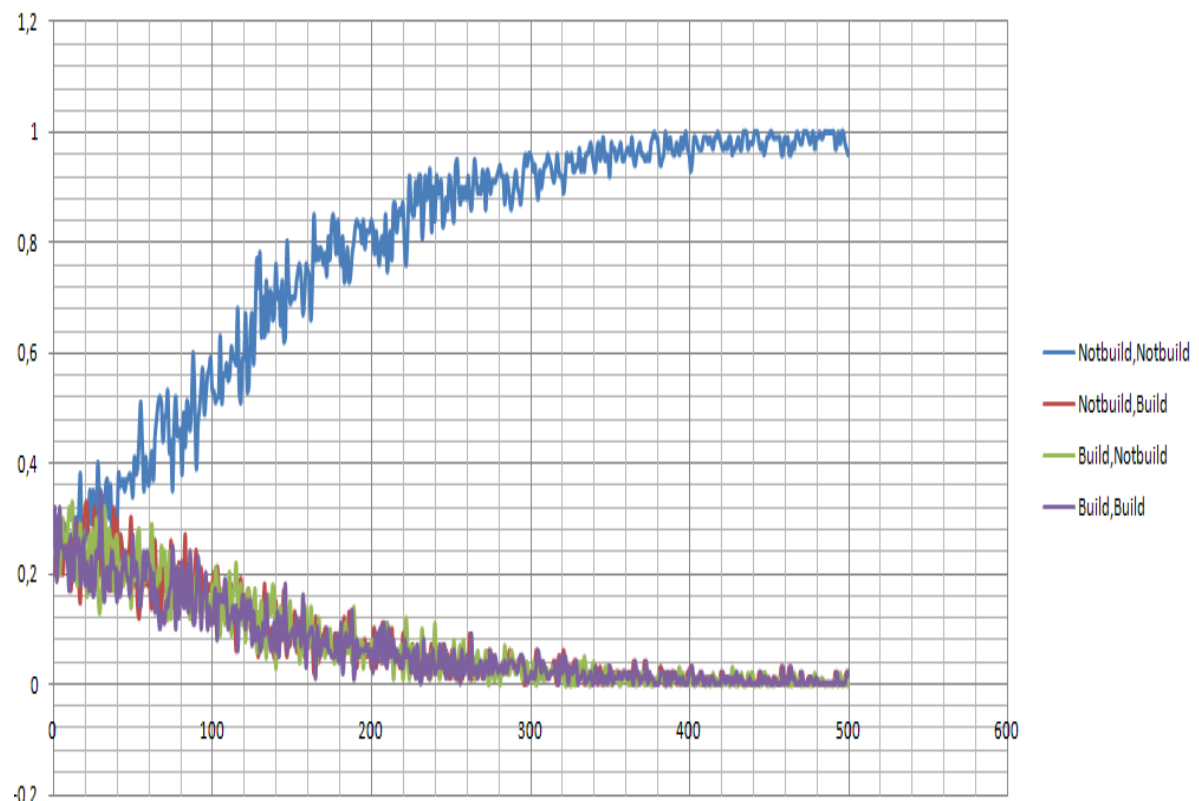


Ο χ άξονας αντιπροσωπεύει γύρους στους οποίους εκπαιδεύεται το σύστημα ενώ ο ψ άξονας αντιπροσωπεύει την πιθανότητα εμφάνισης της κάθε καταστάσεις σε κάθε γύρο. Οι καταστάσεις είναι αυτές που ορίσαμε και στο ποίο πάνω πίνακα. Ο συμβιβασμός συμβολίζεται με την μπλε γραφική, η Ελλάδα εξοπλίζεται συμβολίζεται με την κόκκινη γραφική, η Τουρκία εξοπλίζεται συμβολίζεται με την πράσινη γραφική ενώ η επίθεση συμβολίζεται με την μωβ γραφική.

Όπως είπαμε πιο πάνω βρισκόμαστε σε περίοδο όπου και η Τουρκία ξεδεύει λίγο περισσότερα σε στρατιωτικό εξοπλισμό από την Ελλάδα. Όπως παρατηρούμε στην πιο πάνω γραφική οι δύο χώρες μπορούν να φτάσουν εύκολα σε ένα συμβιβασμό. Αυτό το καταλαβαίνουμε επειδή αρχικά έχουμε τις καταστάσεις να βρίσκονται στα ίδια επίπεδα

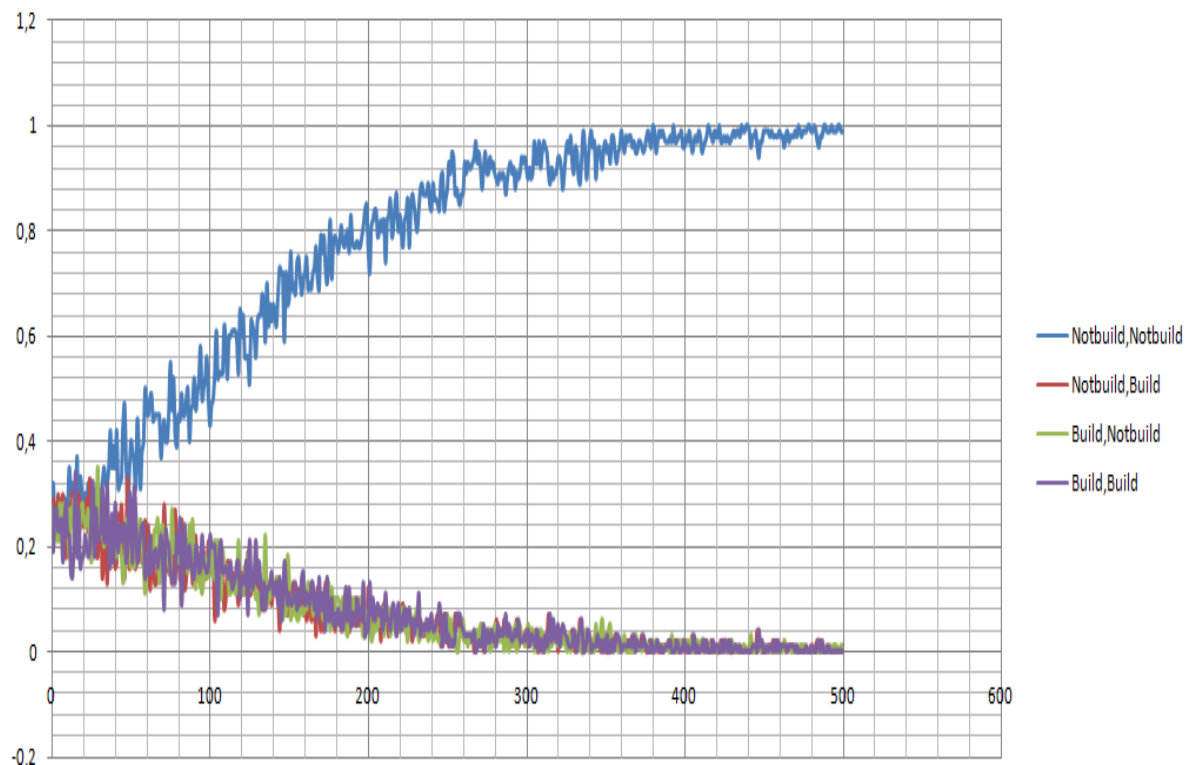
κοντά το 0.2 (από τα 10 που σημαίνει 25% από τα 100) σε πιθανότητα εμφάνισης και σιγά-σιγά βλέπουμε την γραφική του συμβιβασμού (μπλε) να κινείται προς τα πάνω μέχρι την πιθανότητα εμφάνισης 1, ενώ οι υπόλοιπες τρεις γραφικές (που συμβολίζουν τις άλλες τρεις καταστάσεις) κατευθύνονται προς τα κάτω και η πιθανότητα εμφάνισης τους μηδενίζεται. Αυτό σημαίνει ότι σε αυτή την φάση μπορούν εύκολα η δύο χώρες να φτάσουν σε συμβιβασμό. Αυτό είναι λογικό γιατί στην περίοδο 1961-1967 είχαμε την Τουρκία να εξοπλίζεται λίγο πιο πολύ, και επομένως αν οι δαπάνες βρίσκονται στα ίδια επίπεδα ίσως να συμφωνήσουν πιο εύκολα ένα αμοιβαίο συμβιβασμό γλιτώνοντας έτσι περίπου τα ίδια χρήματα και γλιτώνοντας την οικονομία τους παρά σε μια περίοδο όπου και οι δύο εξοπλίζονται ασύστολα για το ποιος θα υπερισχύσει σε στρατιωτικό εξοπλισμό ή σε περίοδο που ο ένας εκ των δύο ξοδεύει πολύ περισσότερα.

Γραφική Παράσταση 4β: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,9 η Τουρκία εξοπλίζεται σε λίγο περισσότερο επίπεδο από ότι η Ελλάδα. Οι τιμές αντικατοπτρίζονται στη χρονιά 1961-1967.



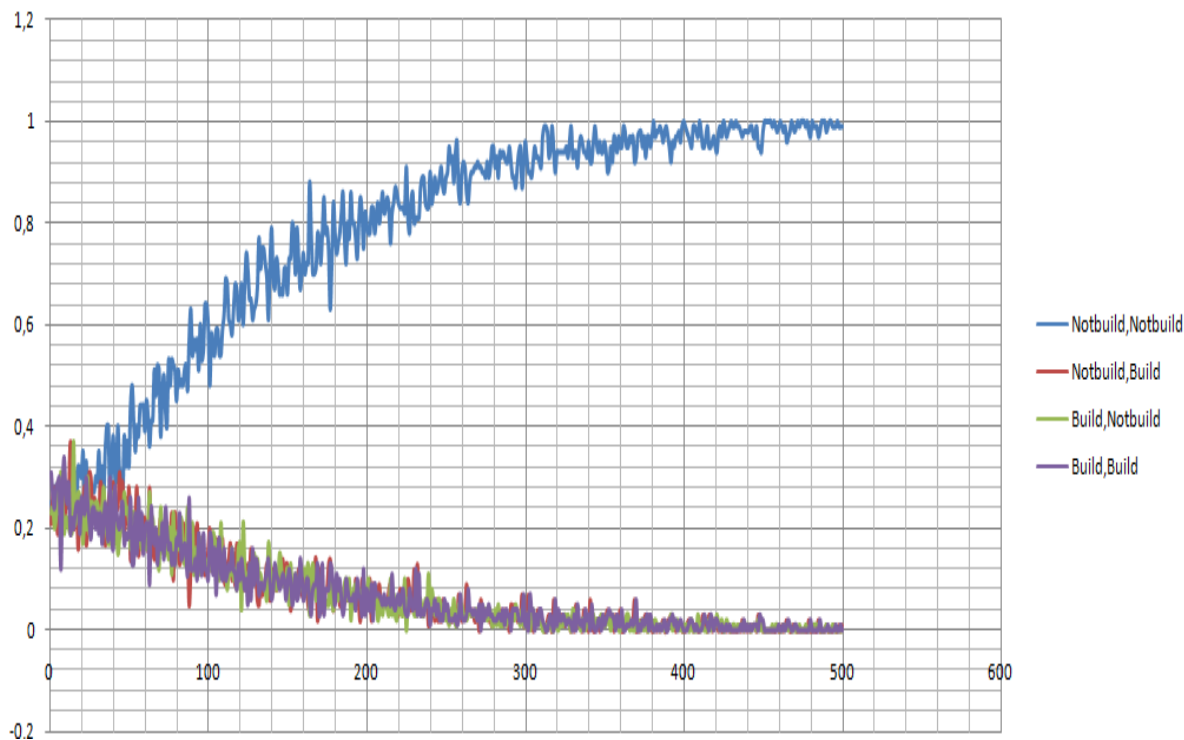
Η μείωση του ρυθμού μάθησης δεν προκάλεσε κάποια αισθητή διαφορά στα αποτελέσματα όπως βλέπουμε από την γραφική. Είναι περίπου τα ίδια με της προηγούμενη βρισκόμαστε σε περίοδο όπου η Τουρκία ξοδεύει λίγο περισσότερα για εξοπλισμό από την Ελλάδα, έτσι ο συμβιβασμός μπορεί να επιτευχθεί πιο εύκολα.

Γραφική Παράσταση 4γ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 ρυθμό έκπτωσης=0,7 η Τουρκία εξοπλίζεται σε λίγο περισσότερο επίπεδο από ότι η Ελλάδα. Οι τιμές αντικατοπτρίζονται στη χρονιά 1961-1967.



Από ότι βλέπουμε έχουμε πάλι μια παρόμοια γραφική με τις προηγούμενες δύο. Η μείωση του ρυθμού έκπτωσης δεν επηρέασε τα αποτελέσματα όμως δεν είναι σωστό να τον μειώνουμε γιατί έτσι οι πράκτορες θα βλέπουν ‘μυωπικά’ θα τους ενδιαφέρουν μόνο η άμεσες ανταμοιβές. Επομένως καλό θα ήταν να έχουμε αρκετά ψηλό ρυθμό έκπτωσης. Σε αυτή την περίπτωση που η Τουρκία ξεοδεύει λίγα περισσότερα από την Ελλάδα είναι ευκολότερο να φτάσουν σε συμβιβασμό οι δύο χώρες.

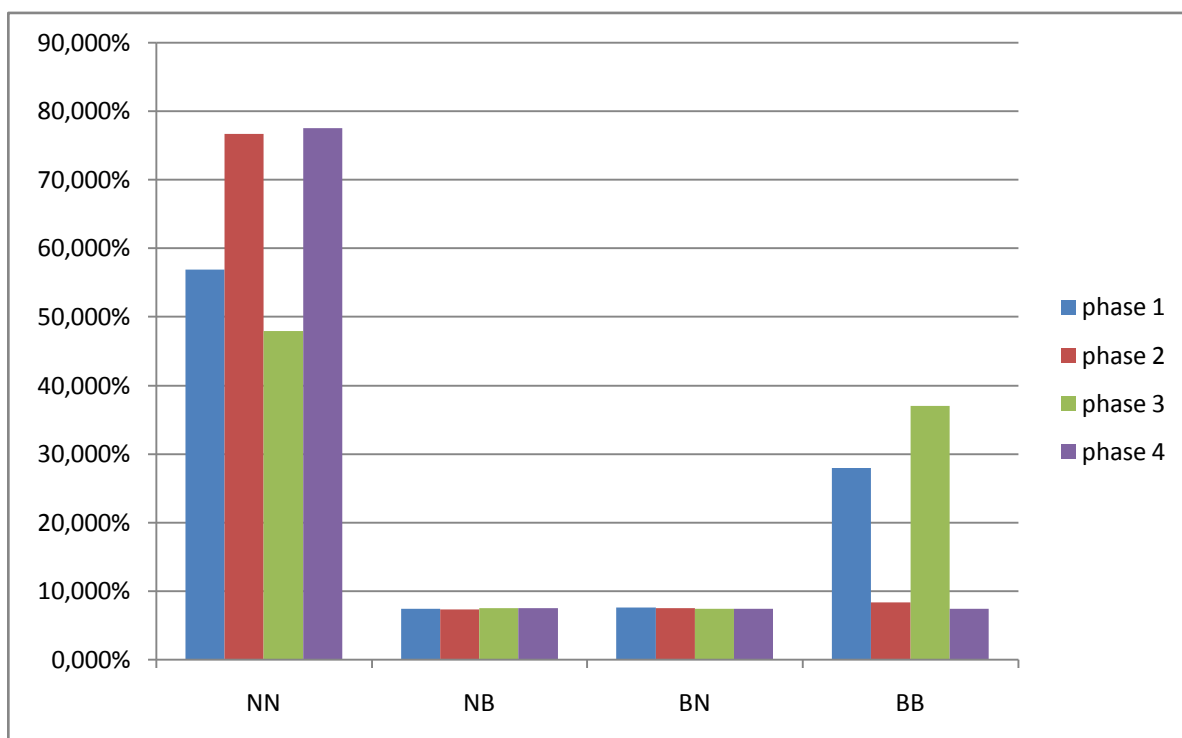
Γραφική Παράσταση 4δ: Πιθανότητα εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,7 ρυθμό έκπτωσης=0,7 η Τουρκία εξοπλίζεται σε λίγο περισσότερο επίπεδο από ότι η Ελλάδα. Οι τιμές αντικατοπτρίζονται στη χρονιά 1961-1967.



Με την μείωση και τον δύο παραμέτρων δεν βλέπουμε κάποια διαφορά από τις προηγούμενες τρεις περιπτώσεις. Το σύστημα μαθαίνει εύκολα τον συμβιβασμό σε αυτή την περίοδο. Όμως καλό είναι να έχουμε ψηλό ρυθμό μάθησης ούτως ώστε οι πράκτορες να μαθαίνουν να φτάνουν γρήγορα σε συμβιβασμό και ψηλό ρυθμό έκπτωσης για να δίνουν περισσότερη αξία στις μακροπρόθεσμες ανταμοιβές.

4.5 Συνένωση Προηγούμενων Φάσεων

Γραφική Παράσταση 5: Ποσοστό εμφάνισης της κάθε κατάστασης με ρυθμό μάθησης=0,9 και ρυθμό έκπτωσης=0,9. Κάθε φάση παίρνει τις τιμές από τους πίνακες της αντιστοίχισης αμοιβών.



Με αυτή την γραφική μπορούμε πόσο της εκατό εμφανίζεται η κάθε κατάσταση σε κάθε φάση. Οι φάσεις είναι αυτές οι 4 που αναφέραμε προηγουμένως σε αυτό το κεφάλαιο και θα τις εξηγήσουμε αναλυτικά. Αρχίζοντας από την φάση 1, όπως είπαμε και πριν βρισκόμαστε στη περίοδο όπου και οι δύο χώρες κάνουν μεγάλες δαπάνες σε στρατιωτικό εξοπλισμό και όπως φαίνεται και από αυτή την γραφική η κατάσταση του συμβιβασμού εμφανίζεται περίπου στο 55% ενώ η κατάσταση της επίθεσης φτάνει σχεδόν το 30%. Επομένως όπως καταλαβαίνουμε ο συμβιβασμός σε αυτή την περίοδο είναι αρκετά δύσκολος με ποσοστά εμφάνισης τόσο κοντινά.

Επίσης όσο αφορά την φάση 2 αναφερόμαστε πάλι στην περίοδο όπου οι δύο χώρες κάνουν ελάχιστα έξοδα σε στρατιωτικό εξοπλισμό. Όπως βλέπουμε η κατάσταση του συμβιβασμού παίρνει πολύ ψηλό ποσοστό εμφάνισης περίπου 80% ενώ η κατάσταση της επίθεσης πιο λίγο από 10%. Γεγονός που μας επιτρέπει να πούμε ότι σε αυτή την περίοδο (όπως είπαμε και πιο πριν) είναι πολύ εύκολος ο συμβιβασμός μεταξύ Ελλάδας και Τουρκίας.

Η Τρίτη φάση αναφαίρετε στην περίοδο όπου η Ελλάδα ξεπερνά κατά πολύ την Τουρκία σε δαπανηρά έξοδα και όπως παρατηρούμε η κατάσταση του συμβιβασμού δεν είναι καθόλου εύκολο να επιτευχθεί λόγω της ελάχιστης διαφοράς ποσοστού εμφάνισης που έχει με την κατάσταση της επίθεσης.

Και τέλος η αναφερόμαστε στη τέταρτη φάση που η Τουρκία κάνει ελάχιστα περισσότερα έξοδα έναντι της Ελλάδας. Εδώ μπορούμε να πούμε ότι ο συμβιβασμός μπορεί να επιτευχθεί άνετα λόγω του μεγάλου ποσοστού εμφάνισης (περίπου 80%) που καταλαμβάνει η κατάσταση του συμβιβασμού, έναντι στην κατάσταση της επίθεσης (λιγότερα από 10%). Όσο αφορά τις καταστάσεις όπου η μια χώρα γίνεται κορόιδο της άλλης βρίσκονται σε χαμηλά και ίδια επίπεδα (λιγότερα από 10%), μας ενδιαφέρουν περισσότερο οι καταστάσεις του συμβιβασμού (NN) και της επίθεσης (BB).

Σειρά προτεραιότητας για Ελλάδα

Όσο αφορά τις αμοιβές θα ορίσουμε την προτεραιότητα ως φθίνουσα σειρά προτεραιότητας δηλαδή θα ξεκινήσουμε από τη πιο επιθυμητή ανταμοιβή προς τη λιγότερο επιθυμητή:

NN1>NN3>NN4>NN2>NB2>NB4>NB3>NB1>BB4>BN2>BB2>BB3>BB4>BB1>BN3>
BN1

Σειρά προτεραιότητας για Τουρκία

Όσο αφορά τις αμοιβές θα ορίσουμε την προτεραιότητα ως φθίνουσα σειρά προτεραιότητας δηλαδή θα ξεκινήσουμε από το πιο επιθυμητό προς το λιγότερο επιθυμητό:

NN4>NN1>=NN2>=BN3>=BN2>=NN3>BN1>BN4>BB3>BB1>=NB2>=BB4>BB2>NB
3>NB1>NB4

Κεφάλαιο 5

Συμπεράσματα και Μελλοντική Εργασία

5.1 Σύνοψη	70
5.2 Συμπεράσματα	73
5.3 Σύγκριση με άλλες μελέτες	75
5.4 Μελλοντική Εργασία	76

5.1 Σύνοψη

Η ΕΜ μέρος της μέρος της Τεχνητής νοημοσύνης έχει ως κύριο σκοπό να εκπαιδεύσει κάποιο πράκτορα αξιολογώντας μια ενέργεια που θα επιλέξει σε κάποια δεδομένη κατάσταση. Αυτή η ενέργεια μπορεί να αξιολογηθεί θετικά δίνοντας την λεγόμενη επιβράβευση στο πράκτορα είτε αρνητικά δίνοντας του τιμωρία. Σύμφωνα με αυτές τις ανταμοιβές θα επιλέξει τις ενέργειες που του δίνουν περισσότερη επιβράβευση ενώ θα απορρίψει επιλογές που του φέρνουν τιμωρίες. Σημαντικό να τονίσουμε ότι ο πράκτορας δεν γνωρίζει από την αρχή ποιες ενέργειες να επιλέξει όμως η ΕΜ παίρνει το λόγο του κριτή και μέσα από την κριτική που του γίνεται με τις αμοιβές και τις τιμωρίες ωθεί τον πράκτορα να ανακαλύψει ποιες ενέργεια θα του επιφέρουν το μεγαλύτερο δυνατό αποτέλεσμα, δηλαδή την μέγιστη ανταμοιβή. Σημαντικό μέρος την ΕΜ είναι και η τακτική την οποία χρησιμοποιεί ο πράκτορας. Έχουμε δύο πολιτικές την άπληστη η οποία εκμεταλλεύεται την υπάρχουσα γνώση και του δίνει άμεση ανταμοιβή και την μη-άπληστη πολιτική που του δίνει την δυνατότητα να εξερευνήσει νέες ενέργειες, που ίσως να του δώσουν και μεγαλύτερη μακροπρόθεσμη ανταμοιβή.

Σε προηγούμενο στάδιο υπήρξε αρκετή μελέτη όσο αφορά την EM σε περιβάλλον ενός πράκτορα, όμως τα τελευταία χρόνια υπήρξε αρκετό ενδιαφέρον για ΕΜΠΠ. Πολλοί πράκτορες εκπαιδεύονται ταυτόχρονα προσπαθώντας να πάρουν την μέγιστη συνολική ενίσχυση. Με αυτό τον τρόπο το περιβάλλον γίνεται δυναμικό.

Η θεωρία παιγνίων βασίζεται σε καταστάσεις που έχουν να κάνουν με τον ανταγωνισμό και προσθέτοντας EM μπορούμε να εκπαιδεύσουμε τους πράκτορες να φτάσουν σε αμοιβαίο συμβιβασμό. Ο κάθε παίχτης ξεχωριστά προσπαθεί να προβλέψει τις επιλογές του αντιπάλου του και να αντιδράσει με τον καλύτερο δυνατό τρόπο προς όφελος του.

Έχουμε παίγνια μηδενικού αθροίσματος στα οποία υπάρχει μόνο ένας νικητής ή γενικά οι επιβραβεύσεις ή οι τιμωρίες κάποιου παίχτη είναι απόλυτα ισορροπημένες με τις επιβραβεύσεις ή τιμωρίες του αντίπαλου παίχτη. Όμως υπάρχουν και τα παίγνια γενικού αθροίσματος που δεν έχουν νικητή ή ηττημένο και με αυτό το είδος παιγνίου έχουμε ασχοληθεί σε αυτή την διπλωματική εργασία.

Μέρος του παιγνίου γενικού αθροίσματος είναι το «Δίλημμα του Φυλακισμένου» στο οποίο όπως αποδεικνύεται η καλύτερη λύση για το κοινό καλό είναι ο συμβιβασμός. Επίσης το «Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου» παίζεται περισσότερο από δύο φορές και είναι χρήσιμο γιατί μοντελοποιεί την εξέλιξη της ανθρώπινης συνεργασίας.

Στόχος αυτής της διπλωματικής εργασίας ήταν να απαντήσει στο ερώτημα αν μπορούμε να μοντελοποιήσουμε το πρόβλημα της κούρσας εξοπλισμού όπλων Ελλάδας Τουρκίας με EM πολλαπλών πρακτόρων και έχοντας ως κύρια βάση το «Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου». Για την EM χρησιμοποιήθηκε ο αλγόριθμος q-learning, που μπορεί να χρησιμοποιηθεί για να βρει μια βέλτιστη πολιτική επιλογή ενέργειας για οποιαδήποτε μακροβιανή διαδικασία απόφασης (MDP). Λειτουργεί μαθαίνοντας συνάρτηση αξίας η οποία δίνει τελικά την αναμενόμενη αξία μιας συγκεκριμένης ενέργειας σε μια δεδομένη

κατάσταση ακολουθώντας την βέλτιστη πολιτική. Όταν μια τέτοια συνάρτηση αξίας μαθαίνεται, η βέλτιστη πολιτική κατασκευάζεται από την απλή επιλογή της ενέργειας με την ψηλότερη αξία σε κάθε κατάσταση. Επιπρόσθετα έχουμε κάνει αντιστοίχιση αμοιβών στο ΔΦ βασιζόμενοι σε ιστορικά γεγονότα που πήραμε από ένα άρθρο. Είχαμε τέσσερις καταστάσεις για μελέτη. Ως πρώτη φάση είχαμε και τις δύο χώρες σε μια περίοδο που εξοπλίζονταν ασύστολα κάνοντας μεγάλες δαπάνες σε στρατιωτικό εξοπλισμό, ως δεύτερη φάση είχαμε και τις δύο χώρες να εξοπλίζονται σε πολύ χαμηλά επίπεδα, ως Τρίτη φάση είχαμε την Ελλάδα να εξοπλίζεται πολύ περισσότερο από την Τουρκία και ως τέταρτη φάση είχαμε την Τουρκία να εξοπλίζεται λίγο περισσότερο από την Ελλάδα. Με βάση αυτά τα σενάρια προέκυψαν οι γραφικές που είδαμε στο κεφάλαιο 4 και βασιζόμενοι σε αυτές βγάλαμε κάποια συμπεράσματα που θα ειπωθούν στο επόμενο κεφάλαιο.

Καλό είναι να λάβουμε υπόψη μας και την οικονομική κατάσταση κάθε χώρας. Για παράδειγμα εάν η μια χώρα εκ των δύο έχει αρκετά καλή οικονομική κατάσταση και κάνει μεγάλες σπατάλες χρημάτων σίγουρα θα είναι αρκετά δύσκολο να την πείσεις να συμβιβαστεί. Αν είμαστε στην πρώτη φάση όπου και οι δύο χώρες σπαταλούν μεγάλα χρηματικά ποσά σε στρατιωτικό εξοπλισμό, τότε αν η μια χώρα ή και οι δύο χώρες έχουν πολύ καλή οικονομική κατάσταση (υψηλό ΑΕΠ) τότε ο συμβιβασμός θα καταστεί ακόμα πιο δύσκολος. Αν είμαστε στην δεύτερη φάση που και οι δύο χώρες σπαταλούν πολύ λίγα σε στρατιωτικό εξοπλισμό και η οικονομία τους είναι καλή τότε ίσως αρχίσουν να ξοδεύουν περισσότερα, με αποτέλεσμα και οι δύο χώρες να μπαίνουν σε μια κούρσα εξοπλισμού και ο συμβιβασμός να γίνεται πολύ δύσκολα. Το ίδιο θα συμβεί και στις άλλες δύο φάσεις όπου στην Τρίτη φάση η Ελλάδα ξοδεύει πολύ περισσότερα και στην τέταρτη φάση η Τουρκία ξοδεύει λίγο περισσότερα σε στρατιωτικό εξοπλισμό. Τότε αν και των δύο χωρών η οικονομική τους κατάσταση είναι καλή θα συνεχίζουν να κάνουν δαπάνες και επίσης αν η οικονομία τους γίνει ακόμα καλύτερη τότε θα ξοδεύουν ακόμα περισσότερα με αποτέλεσμα να μην φτάσουν ποτέ σε συμβιβασμό. Αντίθετα αν και οι δύο χώρες μετά από τόσα χρόνια που βρίσκονται σε κούρσα εξοπλισμού βλέπουν την οικονομία τους να δυσχεράνει όλο και περισσότερο τότε είναι πολύ πιο εύκολο και λογικό να πιεστούν οι δύο χώρες στο να φτάσουν σε ένα αμοιβαίο συμβιβασμό.

5.2 Συμπεράσματα

Σύμφωνα με τα αποτελέσματα μας καταλήγουμε στο συμπέρασμα ότι όντως η «κούρσα εξοπλισμού» μεταξύ Ελλάδας και Τουρκίας μπορεί να μοντελοποιηθεί με ΕΜ και με κύρια βάση το «Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου». Αν πάρουμε αυτή την μοντελοποίηση με το ρεαλιστικό τρόπο μπορούμε να πούμε ότι μέσα από την συνομιλία των δύο χωρών η μία χώρα μαθαίνει από την άλλη μέσα από τις κινήσεις που κάνουν. Παίρνουν τιμωρίες και επιβραβεύσεις για τις ενέργειες που επιλέγουν σε κάθε κατάσταση που βρίσκονται και υπάρχει η συνάρτηση αξίας που σε κάθε γύρο βοηθάει την κάθε χώρα να πάρει την βέλτιστη ενέργεια. Με τα αποτελέσματα που έχουμε πάρει όντως μέσα από την μάθηση μπορούν η δύο χώρες να φτάσουν σε συμβιβασμό.

Ο συμβιβασμός μπορεί να επιτευχθεί ανάμεσα στα δύο κράτη ανάλογα με την πιο πάνω προσέγγιση τηρώντας όμως κάποιες προϋποθέσεις. Παρατηρήθηκε και από τα αποτελέσματα που έχουμε σε προηγούμενες γραφικές ότι ο συμβιβασμός έχει άμεση σχέση με τα έξοδα που κάνει η κάθε χώρα για στρατιωτικό εξοπλισμό. Συγκεκριμένα για να έχουμε συμβιβασμό ανάμεσα στα δύο κράτη πρέπει τα χρήματα που ξοδεύουν σε στρατιωτικό εξοπλισμό να είναι σε χαμηλά επίπεδα καθώς επίσης η διαφορά της δαπάνης που γίνεται μεταξύ τους πρέπει να είναι μικρή. Αν συμβαίνει αυτό το σύστημα μαθαίνει πολύ γρήγορα να φτάνει σε συμβιβασμό. Αντίθετα αν και οι δύο σπαταλούν ασύστολα μεγάλα ποσοστά χρημάτων σε στρατιωτική δύναμή ή η μια χώρα εκ των δύο ξοδεύει πολύ περισσότερα από την αντίπαλη, τότε ο συμβιβασμός γίνεται πιο δύσκολα, με αποτέλεσμα το σύστημα μαθαίνει πιο αργά τον συμβιβασμό.

Σίγουρα θα σας φαίνεται παράξενο το ότι όταν η μια χώρα έκτων δύο σπαταλάει πολύ περισσότερα από την άλλη καθίσταται αρκετά δύσκολος ο συμβιβασμός μεταξύ τους. Όμως σύμφωνα με την αντιστοίχιση που κάναμε σε προηγούμενο κεφάλαιο έπρεπε να υπάρξουν οι ανάλογες διαφορές μεταξύ των τιμών που χρησιμοποιήσαμε και στην φάση 3 όπως είδαμε έπρεπε να χρησιμοποιήσω αρκετά μεγάλη διαφορά ώστε να δείξω την

μεγαλύτερη δαπάνη που έκανε η Ελλάδα έναντι της Τουρκίας. Επομένως αυτή η δυσκολία σε συμβιβασμό που φάνηκε στην φάση 3 πιθανώς να οφείλεται σε κάποιους άλλους παράγοντες που έχουν να κάνουν κυρίως με την χώρα που ξοδεύει πολύ περισσότερα όπως για παράδειγμα η οικονομική κατάσταση της χώρας. Αν η οικονομία μιας χώρας είναι αρκετά καλύτερη από την άλλη και ξοδεύει περισσότερα τότε αν δεν υπάρχει πλήγμα στην οικονομία της τόσο πιο δύσκολο είναι να συμφωνήσει σε ένα συμβιβασμό.

Επίσης πιθανόν και άλλες προϋποθέσεις θα μπορούσαν να παίξουν σημαντικό ρόλο στην εμφάνιση συμβιβασμού σε αυτό το πρόβλημα που δεν έχουν μοντελοποιηθεί σε αυτή την διπλωματική εργασία και ίσως να επιφέρουν και καλύτερα αποτελέσματα. Κάποιες από αυτές θα μπορούσε να ήταν η συμμετοχή έμμεσων παιχτών στο παίγνιο, για παράδειγμα η Αμερική ή Ρωσία. Μιλάμε για χώρες όπου η δύναμη τους υπερσχύει σε μεγάλο βαθμό τη δύναμη Ελλάδας και Τουρκίας ώστε να μπορεί να επιβάλει κάποια τιμωρία σε μια ή και στις δύο χώρες ώστε να αποφασίσουν να συμβιβαστούν. Επίσης η Αμερική θα έπαιζε σημαντικό ρόλο στις κινήσεις των δύο χωρών αν εντασσόταν ως έμμεσος παίχτης στην κούρσα εξοπλισμού διότι εξοπλίζει την Τουρκία και γενικώς ασκεί αρκετή επιρροή. Γεγονός που κάνει τις δύο χώρες να νιώθουν ανασφάλεια χωρίς την βοήθεια της. Δίνοντας τους λοιπόν η Αμερική μια τιμωρία θα τους έκαναν να σκεφτούν πολύ σοβαρά την επιλογή του συμβιβασμού.

Σίγουρα κάποιοι θα σκεφτούνε κατά πόσο αυτό το εργαλείο (η υλοποίηση αυτής της διπλωματικής εργασίας) μπορεί να φανεί χρήσιμο σε πραγματικό επίπεδο. Ας σκεφτούμε το σενάριο ότι η Ελλάδα και η Τουρκία κάνουν μια συνομιλία για να δουν πόσες δαπάνες κάνουν σε στρατιωτικό εξοπλισμό και να δουν κατά πόσο μπορούν να φτάσουν εύκολα σε ένα συμβιβασμό. Αν ένας εκ των δύο μπλοφάρει και λέει ότι θέλει να συμβιβαστεί ώστε να πείσει και τον αντίπαλο να συμβιβαστεί, τότε χρησιμοποιώντας αυτό το εργαλείο μπορεί να καταλάβει ότι εδώ υπάρχει μια μπλόφα. Δηλαδή μπορεί να ισχυρίζεται η μια χώρα ότι θα επιλέξει να συμβιβαστεί όμως το εργαλείο βάζοντας μέσα τα κατάλληλα δεδομένα μπορεί να δείξει ότι ο συμβιβασμός κάτω από κάποιες συνθήκες είναι δύσκολο να επιτευχθεί.

5.3 Σύγκριση με άλλες μελέτες

Σε προηγούμενες εργασίες η προσέγγιση του προβλήματος της «Κούρσας Εξοπλισμού» γινόταν καθαρά με το «Δίλημμα του φυλακισμένου» βασιζόμενο σε κάποιες στρατηγικές, όπως την στρατηγική ‘tit-for-tat’ ή την μικτή στρατηγική. Σε αυτή την διπλωματική εργασία το διαφορετικό που κάνουμε είναι να μοντελοποιήσουμε το πρόβλημα της «Κούρσας Εξοπλισμού» με το παίγνιο γενικού αθροίσματος «Επαναλαμβανόμενο Δίλημμα του φυλακισμένου» αλλά και να προσθέτουμε μέσα EM για εκπαίδευση των πρακτόρων.

Η σημαντικότερη δουλεία που έγινε σε αυτή την διπλωματική εργασία ήταν η αντιστοίχιση των αμοιβών στο «Επαναλαμβανόμενο Δίλημμα του Φυλακισμένου». Μέσα από ιστορικά δεδομένα και με πολύ πειραματισμό βρήκαμε αρκετά βέλτιστες τιμές που να μας αποδίδουν το αποτέλεσμα του συμβιβασμού που θέλουμε εντός λογικών ορίων. Κάθε ενέργεια που επιλέγεται αξιολογείται μέσω της συνάρτησης αξίας που δίνεται από τον αλγόριθμο q-learning. Επίσης χρησιμοποιείται μηχανισμός εξερεύνησης ε-greedy για την εύρεση την βέλτιστης μακροπρόθεσμης συνολικής ανταμοιβής.

5.4 Μελλοντική Εργασία

Αυτή η διπλωματική εργασία θα μπορούσε να επεκταθεί με διάφορους τρόπους. Θα μπορούσε κάποιος να προσθέσει περισσότερο από δύο παίκτες έμμεσα έτσι θα προκύψουν και περισσότερες καταστάσεις για να μελετήσει. Μπορεί να υπάρξουν επίσης και διαφορετικές πολιτικές ή να χρησιμοποιηθεί κάποιο άλλο παίγνιο. Ακόμη θα μπορούσε να χρησιμοποιηθεί κάποιος άλλος αλγόριθμος εκπαίδευσης εκτός από τον Q-learning που χρησιμοποιήσαμε σε αυτή την διπλωματική εργασία. Ένας αλγόριθμος θα μπορούσε να ήταν ο SARSA (State-Action-Reward-State-Action). Θα μπορούσε ακόμη να επεκταθεί το σύστημα έτσι ώστε να συμπεριλάβει μέσα και κάποιους έμμεσους παίκτες (Μεγάλες Δυνάμεις), που σίγουρα θα αποτελέσουν σημαντικό ρόλο στις λήψεις αποφάσεων για κάποιες ενέργειες. Επίσης θα αλλάξουν οι τιμές στις ανταμοιβές και τις τιμωρίες όταν θα υπάρχει κάποιος έμμεσος παίκτης ο οποίος θα τις επηρεάζει. Για παράδειγμα μπορεί να έρθει η Ρωσία σε μια περίοδο όπου η μια χώρα ξοδεύει μεγάλα ποσοστά χρημάτων σε στρατιωτικό εξοπλισμό και να της βάλει μια αρκετά μεγάλη τιμωρία ώστε να σταματήσει τις τόσο μεγάλες δαπάνες και με αυτό τον τρόπο να φτάσουν πιο εύκολα σε συμβιβασμό.

Επίσης αυτή η διπλωματική εργασία θα μπορούσε να χρησιμοποιηθεί για να μοντελοποιήσει κούρσες εξοπλισμών και μεταξύ διαφορετικών χωρών εκτός Ελλάδας και Τουρκίας καθώς και την συμμετοχή περισσότερο από δύο παίκτες [13]. Σήμερα διάφορες χώρες της Βόρειας Ευρώπης (Σουηδία, Πολωνία κ.α) βρίσκονται σε δαπανηρές κούρσες εξοπλισμών λόγω της πολύ επεκταμένης σε στρατιωτικό εξοπλισμό Ρωσίας. Θα μπορούσε αυτή η διπλωματική να επεκταθεί ώστε να καλύψει τους πολλαπλούς πράκτορες που προκύπτουν πιο πάνω για να μπορέσει να δώσει την λύση του συμβιβασμού και να σώσει την οικονομία της κάθε χώρας και ειδικότερα των μικρότερων χωρών που αγοράζουν στρατιωτικό εξοπλισμό για λόγους άμυνας.

Βιβλιογραφία

1. Sutton, R. S. and Barto, A. G. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.
2. <http://129.237.66.221/Sutton-99-MITECS.pdf> 29/05/2014
3. http://www.tu-chemnitz.de/informatik/KI/scripts/ws0910/ml09_5.pdf 29/05/2014
4. http://www.tu-chemnitz.de/informatik/KI/scripts/ws0910/ml09_6.pdf 29/05/2014
5. <http://webdocs.cs.ualberta.ca/~sutton/book/ebook/node66.html> 29/05/2014
6. <http://www.inf.ed.ac.uk/teaching/courses/rl/slides/4rllect10.pdf> 29/05/2014
7. http://www.ias.informatik.tu-darmstadt.de/uploads/Teaching/AutonomousLearningSystems/Kunz_ALS_2013.pdf 29/05/2014
8. <http://webdocs.cs.ualberta.ca/~sutton/book/ebook/node72.html> 29/05/2014
9. <http://pericles.ee.duth.gr/courses/2012-13/AlgCom/DraftSlides/LecAP02-Game%20Theory%20Basics.pdf> 29/05/2014
10. <http://plato.stanford.edu/entries/prisoner-dilemma/#MulMov> 29/05/2014
11. http://sei.pku.edu.cn/~jwp/courses/MultiAgentTechnology/references/Littman_friend-or-foe.pdf 29/05/2014
12. Littman M. L., Markov games as a framework for multi-agent reinforcement learning In Proceedings of the Eleventh International Conference on Machine Learning, pp. 157-163, Morgan Kaufmann, San Francisco, CA, 1994.
13. http://www.math.umass.edu/~lr7q/ps_files/teaching/math456/lecture3.pdf 29/05/2014

14. Ron Smith, Martin Sola, Fabio Spangolo "The Prisoner's Dilemma and Regime-Switching in the Greek-Turkish Arm Race", Journal of Peace Research vol.37, pp.737-750, 2000
15. Stephen J. Majeski "ARMS RACES AS ITERATED PRISONER'S DILEMMA GAMES" Mathematical Social Science 7, vol.7, pp.253-266, 1984
16. Vassilis Vassiliades, Aristodemos Cleanthous, Chris Christodoulou "Multiagent Reinforcement Learning: Spiking and Nonspiking Agents in the Iterated Prisoner's Dilemma", IEEE Transactions on Neural Networks vol.22, pp.639-653, 2011
17. Hu J. and Wellman M.P., Multiagent reinforcement learning: Theoretical framework and an algorithm, In Proceedings of the Fifteenth International Conference on Machine Learning, pp.242-250, Morgan Kaufman, San Francisco, 1998
18. Watkins C.J.C.H, Learning from Delayed Rewards. Ph.D Thesis, Cambridge University, 1989.
19. Bellman R.E., Dynamic Programming. Princeton University Press, Princeton, New Jersey, 1957
20. Nash J.F., Equilibrium Points in N-person Games. In Proceedings of the National Academy of Sciences of the United States of America 36, pp. 48-49, 1950.
21. Axelrod R. and Hamilton W.D., The Evolution of Cooperation. Science, 211:1390-1396, 1981.

Read me file

Για την υλοποίηση του κώδικα σε αυτή την διπλωματική εργασία έχω χρησιμοποιήσει την γλώσσα προγραμματισμού java συγκεκριμένα έχω δουλέψει σε Eclipse Kepler 4.3. Να πούμε επίσης ότι η main βρίσκεται στη κλάση simulation, επομένως αυτή την κλάση πρέπει να τρέξετε, αν θέλετε να δείτε πως λειτουργά το πρόγραμμα.

Παράρτημα Α: Πηγαίος κώδικας

Κλάση Agent

```
import java.util.Random;

public class Agent {

    private static int [][] R=new int[][]{
        {4,4}, //<NotBuild,NotBuild>
        {1,3}, //<NotBuild,Build>
        {3,1}, //<Build,NotBuild>
        {2,2}}; //<Build,Build>

    private static double q[][] = new double[4][2];
    private static double a=0.9;
    private static double GAMMA=0.9;
    double epsilon = 0.0;
    static Random rand = new Random();

    static void initialize()
    {
        for(int i = 0; i < 4; i++)
        {
            for(int j = 0; j < 2; j++)
            {
                q[i][j] = 0;
            } // j
        } // i

        return;
    }

    int chooseAction(int state , double epsilon){
        int action;
        if (rand.nextFloat() < epsilon)
            return action = rand.nextInt(2);
        else
            return greedyAction(state);
    }
}
```

```

int greedyAction(int state)
{
    if ( $q[state][0] > q[state][1]$ ) /* action 0 has greater q value for this state */
        return 0;
    else if ( $q[state][1] > q[state][0]$ ) /* action 1 has greater q value for this state */
        return 1;
    else /* the q values for both actions are equal so choose randomly */
        return rand.nextInt(2);
}

int setReward(int a, int next_state){
int reward= $R[next\_state][a]$ ;
return reward;
}

public void updateValues(int state,int act,int nextState, int r){

     $q[state][act] = (1.0-a)*q[state][act] + a * (r + GAMMA * maximum(nextState))$ ;
    System.out.print("\tq-value " +  $q[state][act]$  + "\t");

}

//Get maximum Q value of this next state based on all possible actions
private static double maximum(final int State)
{
    int winner;
    if (  $q[State][0] > q[State][1]$  )
        winner = 0;
    else if (  $q[State][1] > q[State][0]$  )
        winner = 1;
    else
        winner = rand.nextInt(2);

    return  $q[State][winner]$ ;
}
}

```

Κλάση simulation

```
import java.util.Random;

public class simulation {

    static Agent Turkey=new Agent();
    static Agent Greece=new Agent();

    static int numberOfRuns=50;
    static int numberOfRounds=500;

    static Random rand = new Random();

    static double percentageNN;
    static double percentageNB;
    static double percentageBN;
    static double percentageBB;

    static int counterNN=0;
    static int counterNB=0;
    static int counterBN=0;
    static int counterBB=0;

    static double epsilon;
    static double decade;

    //vectors for each state
    static int[] vectorNN=new int[numberOfRounds];
    static int[] vectorNB=new int[numberOfRounds];
    static int[] vectorBN=new int[numberOfRounds];
    static int[] vectorBB=new int[numberOfRounds];

    //return the current state
    static int getState(int a1, int a2){
        //state 0 nn
        int state=0;
        //00 nn
        if(a1==0 && a2==0)
            state=0;

        //01 nb
        else if(a1==0 && a2==1)
            state=1;
```

```

//10 bn
else if(a1==1 && a2==0)
    state=2;

//11 bb
else if(a1==1 && a2==1)
    state=3;
return state;
}

static void train(int state){
    int action1;
    int action2;

    int nextState;
    int rewardTurkey;
    int rewardGreece;

    while(i<numberOfRounds)    {

        System.out.print("\nround:" + i + "\t");

        action1=Turkey.chooseAction(state ,epsilon);
        action2=Greece.chooseAction(state ,epsilon);

        //immediate reward
        rewardTurkey=Turkey.setReward(action1,state);
        rewardGreece=Greece.setReward(action2,state);

        nextState=getState(action1,action2);

        //update values
        //Turkey
        System.out.print("Turkey ");
        System.out.print(" state " + state);
        System.out.print(" action " + action1);
        Turkey.updateValues(state, action1, nextState, rewardTurkey);

        //Greece
        System.out.print("Greece ");
        System.out.print(" state " + state);
        System.out.print(" action " + action2);
        Greece.updateValues(state, action2, nextState, rewardGreece);
    }
}

```

```

try{
    if(state==0){
        counterNN++;
        if(vectorNN[i]==0){
            vectorNN[i]=1;
            System.out.print("state0->NN( "+vectorNN[i]+" ) ");
        }

        else{
            vectorNN[i]=vectorNN[i]+1;
            System.out.print("state0->NN( "+vectorNN[i]+" ) ");
        }

    }
    //01 nb
    else if(state==1){
        counterNB++;
        if(vectorNB[i]==0){
            vectorNB[i]=1;
            System.out.print("state1->NB( "+vectorNB[i]+" ) ");
        }

        else{
            vectorNB[i]=vectorNB[i]+1;
            System.out.print("state1->NB( "+vectorNB[i]+" ) ");
        }

    }
    //10 bn
    else if(state==2){
        counterBN++;
        if(vectorBN[i]==0){
            vectorBN[i]=1;
            System.out.print("state2->BN( "+vectorBN[i]+" ) ");
        }

        else{
            vectorBN[i]=vectorBN[i]+1;
            System.out.print("state2->BN( "+vectorBN[i]+" ) ");
        }

    }
    //11 bb
    else if(state==3){
        counterBB++;
        if(vectorBB[i]==0){
            vectorBB[i]=1;

```

```

        System.out.print("state3->BB( "+vectorBB[i]+" ) ");
    }
    else{
        vectorBB[i]=vectorBB[i]+1;
        System.out.print("state3->BB( "+vectorBB[i]+" ) ");
    }
}

}

} catch(ArrayIndexOutOfBoundsException e){};

state=nextState;
i++;

}}

public static void simulate(){
    int act1;
    int act2;

    int j=0;
    int currentstate;

    Turkey.initialize(max_reward);
    Greece.initialize(max_reward);

    while(j<numberOfRuns){

        System.out.print("\nrun\n"+j);

        currentstate = rand.nextInt(4);

        train(currentstate);

        j++;

    }
}

```

```

public static void main(String[] args)
{

    simulate();

    System.out.print("\nNN\tNB\tBN\tBB");
    for(int k=0; k<numberOfRuns; k++){
        //System.out.print("\nRound"+k);
        System.out.print("\n");
        System.out.print(vectorNN[k]/(double)numberOfRuns);
        System.out.print("\t"+vectorNB[k]/(double)numberOfRuns);
        System.out.print("\t"+vectorBN[k]/(double)numberOfRuns);
        System.out.print("\t"+vectorBB[k]/(double)numberOfRuns);
    }

    percentageNN=counterNN*100.0/(numberOfRuns*numberOfRuns);
    percentageNB=counterNB*100.0/(numberOfRuns*numberOfRuns);
    percentageBN=counterBN*100.0/(numberOfRuns*numberOfRuns);
    percentageBB=counterBB*100.0/(numberOfRuns*numberOfRuns);

    System.out.print("\nincidence\n"+percentageNN+"%\t"+percentageNB+"%\t"+percentageBN+"%\t"+percentageBB+"%");
    return;
}

}

```