

Ατομική Διπλωματική Εργασία

**ΤΟ ΚΥΠΡΙΑΚΟ ΠΡΟΒΛΗΜΑ ΜΕ ΤΗ ΧΡΗΣΗ ΕΝΙΣΧΥΤΙΚΗΣ
ΜΑΘΗΣΗΣ ΚΑΙ BARGAINING GAME**

Ανδρέας Γιωρκάτζη

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Το κυπριακό πρόβλημα με τη χρήση Ενισχυτικής Μάθησης και Bargaining Game

Ανδρέας Γιωρκάτζη

Επιβλέπων Καθηγητής
Δρ. Χρίστος Χριστοδούλου

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

Μάιος 2022

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, Δρ. Χρίστο Χριστοδούλου, για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τέτοιο θέμα, για τις γνώσεις που μου προσέφερε, τόσο σχετικά με τη Μηχανική Μάθηση όσο και σε πολλούς άλλους τομείς του πεδίου της Πληροφορικής. Η διαρκής προθυμία για να μου προσφέρει βοήθεια και καθοδήγηση και να μου λύσει άμεσα οποιαδήποτε απορία προέκυπτε, ενίσχυσε την προσπάθειά μου για την εκπόνηση αυτής της εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τους γονείς μου αφού χωρίς αυτούς δεν θα ήμουν ο άνθρωπος που είμαι σήμερα και όλη μου την οικογένεια για την ψυχολογική, και όχι μόνο, στήριξη που μου προσέφεραν κατά την διάρκεια των σπουδών μου και των δύσκολων περιόδων της φοίτησής μου.

Τέλος, ευχαριστώ τους φίλους μου για την υποστήριξη, την κατανόηση και την αμέριστη συμπαράσταση που μου παρείχαν.

Περίληψη

Η εργασία αυτή παρουσιάζει μια προσέγγιση του κυπριακού προβλήματος με τη χρήση ενός παιγνίου γενικού αθροίσματος, του bargaining game, και ενισχυτικής μάθησης. Στόχος είναι η αντιστοίχιση ενός πραγματικού, πολύπλοκου προβλήματος σε ένα σύστημα Ενισχυτικής Μάθησης Πολλαπλών Πρακτόρων και η εξαγωγή αποτελεσμάτων για το κατά πόσο αυτό είναι εφικτό και κατά πόσο οι πράκτορες εκπαιδεύονται με στόχο το κοινό καλό. Αποφασίσαμε να ασχοληθούμε με την εδαφική πτυχή του προβλήματος για την οποία υπάρχουν πιο ξεκάθαρες απαιτήσεις. Στο κυπριακό πρόβλημα έχουμε τις 2 άμεσα εμπλεκόμενες πλευρές, τους Ελληνοκύπριους και τους Τουρκοκύπριους, οι οποίοι εδώ και αρκετά χρόνια προσπαθούν να βρουν μια λύση ειρηνικής συμβίωσης. Εκτός όμως από αυτούς υπάρχουν και οι έμμεσα εμπλεκόμενοι στο θέμα, όπως το Ηνωμένο Βασίλειο, ο Οργανισμός Ηνωμένων Εθνών κ.α., οι οποίοι διαδραματίζουν πολύ σημαντικό ρόλο στην έκβαση των διαπραγματεύσεων για τη λύση του προβλήματος. Εξαιτίας της έμμεσης εμπλοκής των εξωτερικών παραγόντων δεν τους χρησιμοποιούμε στο σύστημα ως παίκτες του παιγνίου, αλλά ως Άδηλες Πηγές Ενίσχυσης (ΑΠΕ) οι οποίοι τιμωρούν ή επιβραβεύουν τους παίκτες μας, ο καθένας με γνώμονα τα δικά του συμφέροντα. Σκοπός αυτής της προσθήκης ήταν η εξαγωγή συμπερασμάτων για το αν οι επιλογές των δύο πρακτόρων μπορούν να επηρεαστούν από τους έμμεσα εμπλεκόμενους και αν αυτό έχει την δυνατότητα να συμβάλει στο στόχο της επίλυσης του προβλήματος. Έτσι, καταφέραμε να ενσωματώσουμε περισσότερους από δύο πράκτορες σε παίγνιο δύο πρακτόρων. Αρχικά, η αντιστοίχιση του προβλήματος στο σύστημα έγινε με επιτυχία και οι πράκτορες με τις κατάλληλες παραμέτρους πλησιάζουν όλο και περισσότερο τον συμβιβασμό. Ακόμη, η προσθήκη των ΑΠΕ αύξησε περισσότερο το τελικό ποσοστό συμβιβασμού των δύο πρακτόρων. Αυτό μας δείχνει πως ο ρόλος των εξωτερικών παραγόντων είναι πολύ σημαντικός στην εξέλιξη των διαπραγματεύσεων. Έπειτα, προσθέσαμε και την επιλογή επιπρόσθετης αμοιβής στους πράκτορες, από τους εξωτερικούς παράγοντες, για μια σειρά από ενέργειες των πρακτόρων με βάση το αποτέλεσμα του κάθε γύρου. Αυτή η μέθοδος προσέφερε ακόμα μια αύξηση στα ποσοστά συμβιβασμού, ανεβάζοντας την πιθανότητα συμβιβασμού στο υψηλότερο από όλες τις προσπάθειες, 70%, κάτι το οποίο δείχνει πως με το τέλος ενός γύρου οι διάφορες κινήσεις της κάθε πλευράς μπορούν να επιβραβεύονται ή να τιμωρούνται με σκοπό τη βελτίωση των αποτελεσμάτων επόμενων διαπραγματεύσεων.

Περιεχόμενα

Περίληψη.....	4
Κεφάλαιο 1	9
Εισαγωγή	9
Κεφάλαιο 2	13
Γνωσιολογικό Υπόβαθρο	13
2.1 Στοιχεία Ενισχυτικής Μάθησης	13
2.2 Μέθοδοι Ενισχυτικής Μάθησης	16
2.3 Q-learning algorithm	17
2.4 SARSA algorithm.....	18
2.5 Εξερεύνηση και Αξιοποίηση	19
2.6 Θεωρία Παιγνίων.....	20
2.7 Bargaining Game	21
2.8 Κυπριακό Πρόβλημα.....	23
2.9 Προηγούμενη εργασία.....	25
Κεφάλαιο 3	27
Σχεδιασμός και Υλοποίηση	27

3.1 Εισαγωγή	27
3.2 Σχεδιασμός Συστήματος	29
3.3 Ενσωμάτωση Άδηλων Πηγών Ενίσχυσης	30
3.4 Υλοποίηση	32
Κεφάλαιο 4	37
Αντιστοίχιση Κυπριακού Προβλήματος στο Bargaining game	37
4.1 Εισαγωγή	37
4.2 Περιληπτικά τα κομμάτια του Κυπριακού Προβλήματος που μας ενδιαφέρουν	37
4.3 Αντιστοίχιση του Κυπριακού προβλήματος με το σύστημα	39
4.4 Αντιστοίχιση Εξωτερικών Παραγόντων με το σύστημα	44
Κεφάλαιο 5	53
Αποτελέσματα και Συζήτηση	53
5.1 Εισαγωγή	53
5.2 Προσομοίωση με την χρήση Q-Learning agent	54
5.3 Προσομοίωση με την χρήση SARSA agent	62
5.4 Προσθήκη Άδηλων Πηγών Ενίσχυσης	65

5.5 Προσθήκη Επιπρόσθετης Αμοιβής	69
Κεφάλαιο 6.....	72
Συμπεράσματα και μελλοντική εργασία	72
6.1 Συμπεράσματα.....	72
6.2 Μελλοντική Εργασία.....	74
Παράρτημα Α	79
Παράρτημα Β	80

Λίστα Ακρωνύμιων

EK	Ελληνοκύπριοι/Ελληνοκυπριακή
TK	Τουρκοκύπριοι/Τουρκοκυπριακή
HB	Ηνωμένο Βασίλειο
ΗΠΑ	Ηνωμένες Πολιτείες Αμερικής
ΟΗΕ	Οργανισμός Ηνωμένων Εθνών
ΕΜ	Ενισχυτική Μάθηση
ΕΜΠΠ	Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων
ΧΔ	Χρονική Διαφορά
ΔΠ	Δυναμικός Προγραμματισμός
ΑΠΕ	Άδηλες Πηγές Ενίσχυσης
SARSA	State-Action-Reward-State-Action
ΚΔ	Κυπριακή Δημοκρατία
LR	Learning Rate

Κεφάλαιο 1

Εισαγωγή

Στις μέρες μας, πολλά σημαντικά προβλήματα στα οποία εμπλέκονται δύο πλευρές παραμένουν άλυτα για πολλά χρόνια. Η κάθε πλευρά επιδιώκει να κερδίσει τα μέγιστα για την ίδια χωρίς να λαμβάνει υπόψιν την αντίπαλη πλευρά. Αυτό οδηγεί σε ατέρμονους βρόχους συνομιλιών ή διαδικασιών που δεν επιφέρουν κάποιο αποτέλεσμα.

Η Ενισχυτική Μάθηση(EM) αποτελεί σημαντικό τομέα της Τεχνητής Νοημοσύνης (Kaelbling, Littman, Moore 1996). Είναι μια από τις 3 βασικές μεθόδους Μηχανικής Μάθησης μαζί με την Επιβλεπόμενη και την Μη Επιβλεπόμενη Μάθηση. Σημαντική διαφορά της από την Επιβλεπόμενη μάθηση είναι πως η Ενισχυτική Μάθηση δεν απαιτεί τη γνώση όλων των πιθανών καταστάσεων και δράσεων εκ των προτέρων. Στηρίζεται στη μάθηση μέσω εξερεύνησης, δηλαδή, δεν απαιτεί άμεση ανατροφοδότηση για την κάθε συγκεκριμένη πράξη. Συνεπώς, ο πράκτορας δρα σε ένα προκαθορισμένο περιβάλλον μέσα στο οποίο εκτελεί κάποιες πράξεις για τις οποίες αποφασίζει με δικά του κριτήρια. Ο πράκτορας ανταμείβεται ή τιμωρείται με βάση τις πράξεις τις οποίες ακολούθησε.

Η τακτική με την οποία ο πράκτορας επιλέγει τις πράξεις του, βασίζεται στο κατά πόσο έχει την επιλογή να εξερευνήσει το περιβάλλον του (ή όχι). Άπληστη τακτική ορίζεται η τακτική κατά την οποία ο πράκτορας ακολουθεί εξ ολοκλήρου την ήδη υπάρχουσα γνώση χωρίς να εξερευνά άλλες πιθανές επιλογές. Αντίθετα, μια τακτική ορίζεται μη άπληστη όταν ο πράκτορας δεν αξιοποιεί την γνώση την οποία απέκτησε αλλά συνεχώς εξερευνά το περιβάλλον. Υπάρχουν επίσης τακτικές οι οποίες αποτελούν συνδυασμό των δύο αυτών τακτικών και χρησιμοποιούν τόσο την υπάρχουσα γνώση όσο και την εξερεύνηση του περιβάλλοντος.

Αρχικά, η EM περιοριζόταν σε περιβάλλοντα με ένα πράκτορα, μελλοντικά όμως δημιουργήθηκε η ανάγκη της επέκτασης της και συνεπώς της αφαίρεσης αυτού του περιορισμού με την εισαγωγή περιβαλλόντων με πολλαπλούς πράκτορες (Littman 1994,

Hu & Wellman 1998). Με την εμφάνιση της Ενισχυτικής Μάθησης Πολλαπλών Πρακτόρων όπως ονομάζεται, πολλά πραγματικά προβλήματα είχαν τη δυνατότητα πλέον να μοντελλοποιηθούν με τη χρήση EM.

Ένα χρήσιμο εργαλείο για την πιο ρεαλιστική μοντελλοποίηση των προβλημάτων αυτών είναι η Θεωρία Παιγνίων (Fudenberg & Tirole 1991). Τα παίγνια, τα οποία πρωτοεμφανίστηκαν ως κλάδος των οικονομικών (Neumann & Morgenstern 1944), αποτελούν την ανάλυση ανταγωνιστικών καταστάσεων με μαθηματική ακρίβεια μέσω μαθηματικών μοντέλων. Αυτές οι δυνατότητες των παιγνίων φάνηκαν πολύ χρήσιμες στη μοντελλοποίηση πραγματικών προβλημάτων με τη χρήση του παιγνίου ως το περιβάλλον μιας προσομοίωσης με EM (Nowé Vrancx & Hauwere 2012)

Υπάρχουν δύο είδη παιγνίων που μας ενδιαφέρουν. Τα παίγνια μηδενικού αθροίσματος και τα παίγνια μη-μηδενικού αθροίσματος. Μηδενικού αθροίσματος ονομάζονται τα παίγνια στα οποία η νίκη/κέρδος του ενός παίκτη (πράκτορα στην περίπτωση μας) αντιστοιχεί στην ήττα/απώλεια του/των αντίπαλου/ων. Αντίθετα στα μη-μηδενικού (γενικού) αθροίσματος μπορούν να υπάρξουν όλοι νικητές/κερδισμένοι.

Το κυπριακό πρόβλημα με το οποίο θα ασχοληθούμε στην συγκεκριμένη διπλωματική εργασία είναι ένα πρόβλημα το οποίο χρονολογείται δεκάδες χρόνια πίσω. Σε αυτό το πρόβλημα υπάρχουν 2 άμεσα εμπλεκόμενοι πράκτορες, η Ελληνοκυπριακή πλευρά (ΕΚ) και η Τουρκοκυπριακή πλευρά (ΤΚ). Υπάρχουν όμως και έμμεσα εμπλεκόμενοι πράκτορες οι οποίοι είναι κοινά αποδεκτό πως επηρεάζουν την έκβαση των διαπραγματεύσεων, όπως είναι η Ελλάδα, η Τουρκία, τα Ηνωμένα Έθνη, το Ηνωμένο Βασίλειο και οι Ηνωμένες Πολιτείες Αμερικής.

Μετά την στρατιωτική εισβολή από την Τουρκία στην Κύπρο το 1974, η Τουρκία παρακρατεί περίπου το 37% του εδάφους της Κυπριακής Δημοκρατίας. Από το 1975 και μετά επανήρχισαν οι διαδικασίες για την εξεύρεση μιας κοινής λύσης που θα ευχαριστεί και τις δύο πλευρές, λύση η οποία μέχρι και σήμερα (2022) δεν έχει βρεθεί. Στη συγκεκριμένη εργασία θα ασχοληθούμε με το εδαφικό κομμάτι του προβλήματος. Ο λόγος για τον οποίο αποφασίσαμε να ασχοληθούμε με την εδαφική πτυχή του προβλήματος είναι διότι αποτελεί μια καλή ένδειξη για το εάν μπορεί να

μοντελλοποιηθεί ένα τόσο πολύπλοκο πρόβλημα και προσφέρεται λόγω των ξεκάθαρων θέσεων (με αριθμούς) και των δυο πλευρών.

Στο πρόβλημα αυτό μπορεί να συνυπάρξουν και δύο νικητές (κερδισμένοι) για αυτό το λόγο θα χρησιμοποιήσουμε παίγνιο μη-μηδενικού αθροίσματος. Οι δύο άμεσα εμπλεκόμενες πλευρές είναι πιθανόν να καταλήξουν και οι 2 κερδισμένες αλλά και οι 2 ηττημένες. Υπάρχει φυσικά και το ενδεχόμενο η μια πλευρά να πάρει αυτό που θέλει ενώ η άλλη όχι.

Το παίγνιο στο οποίο θα βασιστούμε στην συγκεκριμένη διπλωματική είναι το Rubinstein's Bargaining Game (Rubinstein 1982). Είναι ένα απλό παιχνίδι γενικού αθροίσματος κατά το οποίο οι δύο παίκτες έχοντας ένα ποσό προσπαθούν να συμφωνήσουν στη διαμοίραση του ποσού μεταξύ τους. Στο παιχνίδι αυτό υπάρχει επίσης η έννοια της μείωσης του ποσού σε κάθε γύρο (Nash 1950), άρα όσο πιο γρήγορα οι δύο παίκτες φτάσουν σε συμφωνία τόσο καλύτερο μπορεί να είναι το αποτέλεσμα και για τους δύο.

Επιλέξαμε αυτό το παίγνιο λόγω του ότι για την εδαφική πτυχή του προβλήματος έχουμε ένα σταθερό ποσό το οποίο πρέπει να μοιραστούν οι δύο παίκτες και με την πάροδο του χρόνου δημιουργούνται άλλα προβλήματα, όπως αντιπαραθέσεις ή εμπλοκή τρίτων για διεκδίκηση εδάφους με αφορμή την συμφιλίωση, προβλήματα τα οποία στην συγκεκριμένη περίπτωση προσομοιώνονται με την μείωση του ποσού. Το ποσό αυτό που θα μειώνεται συνεχώς το ορίζουμε ως το Pot δηλαδή την ποσότητα που περιέχει το «καλάθι» από το οποίο διεκδικούν οι δύο πράκτορες.

Στη διπλωματική αυτή θα δούμε επίσης την επιρροή που μπορούν να φέρουν οι έμμεσα εμπλεκόμενοι παίκτες σε ένα τόσο περίπλοκο πρόβλημα μέσω των αμοιβών και τιμωριών που μπορούν να δώσουν στους 2 παίκτες. Οι αμοιβές και τιμωρίες είναι στην πραγματικότητα τα ωφέληματα και οι κυρώσεις αντίστοιχα που δίνονται στις 2 πλευρές. Οι εξωτερικοί αυτοί παράγοντες όπως και στην πραγματικότητα δεν είναι άμεσα εμπλεκόμενοι παίκτες αλλά έμμεσα συνεπώς δεν θέλαμε να είναι παίκτες του παίγνιου.

Ο τρόπος με τον οποίο θα εντάξουμε την ιδέα των αμοιβών των εξωτερικών παραγόντων είναι αξιοποιώντας κάποιες γνώσεις από την ψυχολογία. Πιο συγκεκριμένα θα θεωρήσουμε πως οι έμμεσα εμπλεκόμενες χώρες είναι Άδηλες Πηγές Ενίσχυσης (ΑΠΕ) (Γεωργίου, 2007; Deci, 1972; Ryan & Deci, 2000). Οι ΑΠΕ χαρακτηρίζονται ως παράγοντες οι οποίοι με την επιρροή τους στους 2 παίκτες έχουν επιρροή και στο τελικό αποτέλεσμα. Οπότε, οι δύο πλευρές (ΕΚ, ΤΚ) θα θεωρούνται βασικοί παίκτες του παίγνιου ενώ οι υπόλοιποι εμπλεκόμενοι (Ελλάδα, Τουρκία, Ηνωμένο Βασίλειο, Ηνωμένα Έθνη και Ηνωμένες Πολιτείες Αμερικής) υπάρχουν στο σύστημα ως πηγές που επιβραβεύουν ή τιμωρούν τις αποφάσεις των 2 παικτών ανάλογα με τα δικά τους ή και τα κοινά συμφέροντα.

Κεφάλαιο 2

Γνωσιολογικό Υπόβαθρο

2.1 Στοιχεία Ενισχυτικής Μάθησης	13
2.2 Μεθόδους Ενισχυτικής Μάθησης	16
2.3 Q-Learning	17
2.4 SARSA	18
2.5 Θεωρία Παιγνίων	19
2.5 Θεωρία Παιγνίων	20
2.6 Bargaining Game	21
2.7 Κυπριακό Πρόβλημα	23
2.8 Προηγούμενη Εργασία	25

2.1 Στοιχεία Ενισχυτικής Μάθησης

Η Ενισχυτική Μάθηση είναι μια από τις 3 βασικές μεθόδους της Μηχανικής Μάθησης. Είναι η μάθηση κατά την οποία ο εκπαιδευόμενος μαθαίνει μέσα από τις εμπειρίες του. Δεν υπάρχει ο δάσκαλος όπως σε άλλες μορφές μάθησης αλλά υπάρχει η έννοια του κριτή, ο οποίος δίνει αμοιβή ή τιμωρία στον μαθητευόμενο ανάλογα με το αποτέλεσμα της πράξης του. Αυτό σημαίνει πως ο μαθητευόμενος δεν γνωρίζει ποτέ ποια είναι η σωστή απόφαση που πρέπει να πάρει και δεν υπάρχει κάποιος για να του δείξει τη σωστή απάντηση.

Χαρακτηρίζεται ως η πιο φυσιολογική μορφή μάθησης αφού μοιάζει με τον τρόπο που μαθαίνουμε στην καθημερινότητά μας. Υπάρχει ο μαθητευόμενος ο οποίος ονομάζεται πράκτορας και αλληλοεπιδρά με το περιβάλλον στο οποίο βρίσκεται. Μερικές από τις πράξεις του προσφέρουν κάποια επιβράβευση, κάποιες άλλες προσφέρουν κάποια

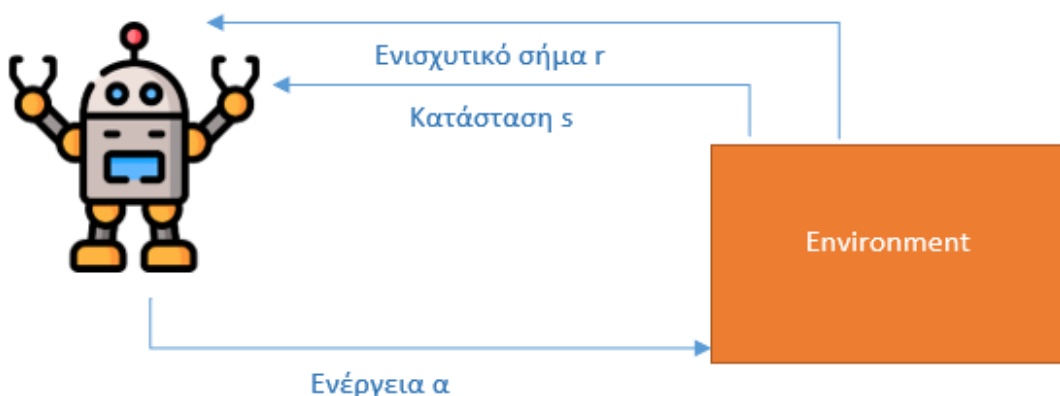
τιμωρία και υπάρχουν και πράξεις οι οποίες μπορεί να μην προσφέρουν τίποτα. Οι πράξεις ή αλλιώς ενέργειες που μπορεί να εκτελέσει ο πράκτορας εξαρτώνται από το περιβάλλον στο οποίο βρίσκεται και από τις δυνατότητες τις οποίες έχει. Σκοπός του πράκτορα είναι να αποκομίσει όσο το δυνατόν περισσότερο κέρδος μπορεί.

Ο πράκτορας εκτελεί μία από τις πράξεις που μπορεί να εκτελέσει ανάλογα με τη συγκεκριμένη κατάσταση στην οποία βρίσκεται. Έπειτα το περιβάλλον του επιστρέφει την αμοιβή ή τιμωρία ανάλογα με την πράξη του, αυτό ορίζεται ως το ενισχυτικό σήμα. Στην συνέχεια η κατάσταση του πράκτορα ενημερώνεται και αυτή η διαδικασία συνεχίζεται.

Η ΕΜ μέσω των πράξεων του κάθε γύρου επιδιώκει τη δημιουργία ενός δικτύου που μπορεί να λάβει υπόψιν μελλοντικές πράξεις οι οποίες ακόμη δεν έχουν εξερευνηθεί. Αυτό μπορεί να καταστεί δυνατό αφού οι ανταμοιβές οι οποίες δίνονται δεν επηρεάζουν μόνο τη συγκεκριμένη κατάσταση στην οποία βρίσκεται ο πράκτορας αλλά και μεταγενέστερες καταστάσεις.

Συνοψίζοντας τα πιο πάνω η Ενισχυτική Μάθηση αποτελείται από 3 βασικά στοιχεία:

- Τον πράκτορα (ο μαθητευόμενος)
- Το περιβάλλον
- Τις πράξεις (ενέργειες) που μπορεί να έχει ο πράκτορας.



Σχήμα 2.1.1

Σχήμα Μοντέλου ενισχυτικής μάθησης ενός πράκτορα

Ο πράκτορας αλληλεπιδρά με το περιβάλλον με την εκτέλεση κάποιων ενεργειών. Το περιβάλλον μεταβάλλει την κατάσταση του πράκτορα και του επιστρέφει το ενισχυτικό σήμα το οποίο μπορεί να είναι επιβράβευση ή τιμωρία με βάση την ενέργεια που ακολούθησε.

Το σχήμα 2.1.1 πιο πάνω παρουσιάζει τα 3 βασικά στοιχεία της ΕΜ και το πώς αυτά αλληλεπιδρούν μεταξύ τους. Συνοπτικά, ο πράκτορας εκτελεί μια πράξη η οποία επιφέρει κάποιες αλλαγές στο περιβάλλον, το περιβάλλον επιστρέφει το ενισχυτικό σήμα (ανταμοιβή ή τιμωρία) στον πράκτορα μαζί με τη νέα κατάσταση στην οποία βρίσκεται πλέον ο πράκτορας.

Η επιλογή της πράξης που θα εκτελέσει ο πράκτορας, ή αλλιώς τακτική, είναι ένα από τα πιο σημαντικά κομμάτια της εκπαίδευσης. Η τακτική καθορίζει με ποια κριτήρια ο πράκτορας θα αποφασίσει ποια πράξη θα εκτελέσει. Είναι μια συνάρτηση η οποία κάνει την αντιστοίχιση της κατάστασης με την πράξη.

Ο πράκτορας έχει την επιλογή να εκμεταλλευτεί την ήδη υπάρχουσα γνώση από προηγούμενες εμπειρίες αλλά και τη δυνατότητα να εξερευνήσει νέες καταστάσεις μέσω πράξεων των οποίων ακόμη το αποτέλεσμα είναι άγνωστο. Αυτές οι δύο τακτικές ονομάζονται άπληστη (greedy) και μη-άπληστη (non-greedy) αντίστοιχα. Υπάρχει βέβαια και η ενδιάμεση λύση η οποία είναι μια πιο ισορροπημένη τακτική μεταξύ των δύο.

Στην ΕΜ όπως έχει προαναφερθεί σκοπός του πράκτορα είναι να μεγιστοποιήσει το κέρδος (reward) μακροπρόθεσμα. Για να επιτευχθεί αυτό η υλοποίηση της ΕΜ απαιτεί δύο συναρτήσεις, τη Συνάρτηση Αμοιβής και τη Συνάρτηση Αξίας.

Η Συνάρτηση Αμοιβής (Reward function) δίνει στον πράκτορα επιβράβευση ανάλογα με το τι είναι καλό βραχυπρόθεσμα. Ουσιαστικά συνδέει την κατάσταση με την πράξη (state-action pair) σε ένα Ενισχυτικό Σήμα (Κριτής).

Η Συνάρτηση Αξίας (Value function) δίνει στον πράκτορα επιβράβευση ανάλογα με το τι είναι καλό μακροπρόθεσμα.

Χρησιμοποιείται επίσης ο ρυθμός έκπτωσης. Ο ρυθμός έκπτωσης προσφέρει στον πράκτορα τη δυνατότητα είτε να «βλέπει» και να παίρνει αποφάσεις με στόχο το μακροπρόθεσμο κέρδος ή το άμεσο κέρδος. Ένας ρυθμός έκπτωσης κοντά στο μηδέν ουσιαστικά δείχνει πώς ο πράκτορας παίρνει αποφάσεις αναλογιζόμενος μόνο το άμεσο κέρδος. Αντίθετα, ένας ρυθμός μάθησης κοντά στο 1 δείχνει πώς ο πράκτορας όταν θα αποφασίσει για την πράξη του θα λάβει υπόψιν του και μελλοντικές αμοιβές.

Στο περιβάλλον όπου κινείται ο πράκτορας μας μπορούν να υπάρχουν και άλλοι πράκτορες. Αυτό ονομάζεται Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων και δίνει τη δυνατότητα να μοντελοποιήσουμε πιο πολύπλοκα και πραγματικά προβλήματα αφού, εκ των πραγμάτων, δεν μπορεί να υπάρξει διαμάχη (conflict) χωρίς να υπάρξουν τουλάχιστον δυο πλευρές.

2.2 Μέθοδοι Ενισχυτικής Μάθησης

Οι μέθοδοι της ΕΜ αποτελούν τον τρόπο με τον οποίο ο πράκτορας δρα και ενημερώνει την πολιτική του με τη χρήση της εμπειρίας του. Υπάρχουν τρεις κύριες μέθοδοι. Ο Δυναμικός προγραμματισμός (ΔΠ) (Bellman, 1957), η Monte Carlo (Metropolis et al, 1949) και η μάθηση Χρονικών Διαφορών (Temporal Difference) (Tesauro, 1991) με την οποία και θα ασχοληθούμε σε αυτή τη διπλωματική. Όλες οι μέθοδοι χρησιμοποιούν την αξιολόγηση πολιτικής (policy evaluation), δηλαδή έχουν τη συνάρτηση αξίας την οποία εκτιμούν συνεχώς με τη χρήση των αξιών των καταστάσεων και επίσης, βελτιώνουν την πολιτική (policy improvement), που χρησιμοποιεί η συνάρτηση αξίας.

Ο Δυναμικός Προγραμματισμός (ΔΠ) αποτελεί μια οικογένεια αλγορίθμων. Οι αλγόριθμοι αυτοί επιδιώκουν να βρουν τη βέλτιστη πολιτική με τη χρήση Συναρτήσεων Αξίας και την αξιοποίηση του μοντέλου περιβάλλοντος και απαιτούν μεγάλη υπολογιστική δύναμη και γνώση του μοντέλου περιβάλλοντος. Δηλαδή, προϋποθέτουν πως γνωρίζουμε τη συνάρτηση μετάβασης καταστάσεων και τη συνάρτηση ανταμοιβών για την κάθε κατάσταση. Αυτό σε πολλές περιπτώσεις όπως και στη δική μας είναι πρακτικά αδύνατο αφού έχουμε ένα δυναμικό περιβάλλον.

Σε αντίθεση με τον ΔΠ οι μέθοδοι Monte Carlo δεν απαιτούν εκ των προτέρων τη γνώση του μοντέλου περιβάλλοντος. Ο τρόπος με τον οποίο μαθαίνουν είναι με βάση την εμπειρία που αποκτά ο πράκτορας κατά την διάρκεια της εξερεύνησης του αφού υπάρχει η δυνατότητα να μαθαίνει από απευθείας (on-line) ή εξομοιούμενη (off-line) αλληλεπίδραση με το περιβάλλον. Με την απευθείας μάθηση δεν είναι απαραίτητη καμία προηγούμενη γνώση σχετικά με το περιβάλλον στο οποίο δρα ο πράκτορας. Για την εξομοιούμενη εμπειρία απαιτείται μια προσέγγιση του περιβάλλοντος και όχι ένα πλήρες μοντέλο. Ο βασικός πυρήνας των μεθόδων Monte Carlo είναι η χρήση των μέσων όρων των απολαβών του πράκτορα.

Η μάθηση χρονικών διαφορών (ΧΔ) συνδυάζει τις δύο προαναφερθείσες μεθόδους. Αρχικά, όπως και στον ΔΠ, για την ενημέρωση των συναρτήσεων χρησιμοποιούν εκτιμήσεις και δεν περιμένουν για το τελικό αποτέλεσμα. Επίσης, όπως και οι μέθοδοι Monte Carlo δεν απαιτούν τη γνώση του μοντέλου περιβάλλοντος αλλά μπορούν να λειτουργήσουν και σε ένα άγνωστο περιβάλλον. Μια βασική μέθοδος ΧΔ είναι η TD(0) για την οποία η συνάρτηση ενημέρωσης (2.2.1) έχει ως εξής:

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (2.2.1)$$

Όπου s_t είναι η κατάσταση τη χρονική στιγμή t , α είναι ο ρυθμός μάθησης και γ ο ρυθμός έκπτωσης. Είναι προφανές από τη συνάρτηση πώς η μέθοδος TD χρησιμοποιεί μια εκτίμηση για μελλοντική κατάσταση $V(s_{t+1})$ έτσι ώστε να ενημερώσει την τρέχουσα αξία. Τα κύρια πλεονεκτήματα των μεθόδων ΧΔ είναι ότι δεν απαιτούν την γνώση του μοντέλου περιβάλλοντος και με την πάροδο του κάθε βήματος μπορεί να γίνεται η ενημέρωση, δεν χρειάζεται να τερματιστεί το επεισόδιο όπως στις μεθόδους Monte Carlo.

2.3 Q-learning algorithm

Ο Αλγόριθμος Q-learning (Watkins, 1989) είναι ένας αλγόριθμος της μεθόδου Χρονικής Διαφοράς. Ο Q-learning υπολογίζει το $Q(s,a)$ (τη συνολική ανταμοιβή με την έκπτωση) ξεκινώντας από την κατάσταση s και εκτελώντας την πράξη a . Χαρακτηρίζεται ως εκτός-τακτικής αλγόριθμος (off-policy algorithm) αφού μπορεί να χρησιμοποιήσει

οποιαδήποτε τακτική για να υπολογίσει το Q ανεξάρτητα από την τακτική με την οποία ο πράκτορας αποφασίζει για τις πράξεις.

Στην περίπτωση του του αλγόριθμου αυτού η συνάρτηση ενέργειας – αξίας ονομάζεται Q και προσεγγίζει την Q^* η οποία αποτελεί την optimal συνάρτηση. Η εξίσωση που χρησιμοποιείται (2.3.1) φαίνεται πιο κάτω και ουσιαστικά υπολογίζει την εκτιμώμενη αξία του ζευγαριού κατάστασης (state) – ενέργειας (action) $\langle s, a \rangle$:

$$Q(s, a) \leftarrow [1 - lr]Q(s, a) + lr[r(s, a) + \gamma \max_b Q(s', b)] \quad (2.3.1)$$

Όπου $Q(s, a)$ η συνάρτηση ενέργειας-αξίας του ζευγαριού $\langle s, a \rangle$, lr ο ρυθμός μάθησης, $r(s, a)$ η αμοιβή που επιστρέφει η πράξη a στην κατάσταση s , γ ο ρυθμός έκπτωσης και $\max_b Q(s', b)$ η συνάρτηση ενέργειας-αξίας του ζευγαριού $\langle s', b \rangle$ το s' είναι η επόμενη κατάσταση και το b η ενέργεια με τη μέγιστη τιμή της συνάρτησης Q .

Η μέθοδος η οποία ακολουθείται παρουσιάζεται στον αλγόριθμο 2.3.2 που φαίνεται πιο κάτω υπό τη μορφή ψευδοκώδικα.

Αλγόριθμος Q-learning 2.3.2

```

Initialize  $Q(s, a)$ 
for each episode do
  observe state  $s$ 
  repeat
    select action  $a$  evaluating  $Q$ 
    take action  $a$ 
    observe  $r, s'$ 
     $Q(s, a) \leftarrow [1 - lr]Q(s, a) + lr[r(s, a) + \gamma \max_b Q(s', b)]$ 
     $s \leftarrow s'$ 
  until  $s$  is terminal
end for

```

2.4 SARSA algorithm

Ο αλγόριθμος SARSA(STATE-ACTION-REWARD-STATE-ACTION) είναι επίσης μια μέθοδος ΧΔ, αλλά αυτός ο αλγόριθμος είναι εντός τακτικής (on-policy).

Αλγόριθμος εντός τακτικής σημαίνει πως χρησιμοποιεί την ίδια τακτική με αυτήν που χρησιμοποιείται για την επιλογή των ενεργειών και τη βελτιώνει συνεχώς (τακτική επιλογής ενέργειας = τακτική στόχου). Η εξίσωση που χρησιμοποιείται φαίνεται πιο κάτω (2.4.1):

$$Q(s, a) \leftarrow Q(s, a) + lr[r + \gamma Q(s', a') - Q(s, a)] \quad (2.4.1)$$

Όπου $Q(s, a)$ η συνάρτηση ενέργειας-αξίας του ζευγαριού $\langle s, a \rangle$, lr ο ρυθμός μάθησης, r η αμοιβή που επιστρέφει αυτή η πράξη, γ ο ρυθμός έκπτωσης και $Q(s', a')$ η συνάρτηση ενέργειας-αξίας του ζευγαριού $\langle s', a' \rangle$ με το s' να είναι η επόμενη κατάσταση και το a' η επόμενη ενέργεια.

Η μέθοδος η οποία ακολουθείται παρουσιάζεται στον αλγόριθμο 2.4.2 που φαίνεται πιο κάτω υπό τη μορφή ψευδοκώδικα.

Αλγόριθμος SARSA 2.4.2

```

Initialize  $Q(s, a)$ 
for each episode do
  observe state  $s$ 
  select action  $a$  evaluating  $Q$ 
  repeat
    take action  $a$ 
    observe  $r, s'$ 
    Select action  $a'$  evaluating  $Q$ 
     $Q(s, a) \leftarrow Q(s, a) + lr[r + \gamma Q(s', a') - Q(s, a)]$ 
     $s \leftarrow s', a \leftarrow a'$ 
  until  $s$  is terminal
end for

```

2.5 Εξερεύνηση και Αξιοποίηση

Στην EM έχουμε επίσης το θέμα της εξερεύνησης (exploration) και της αξιοποίησης (exploitation), δηλαδή αν ο πράκτορας θα αξιοποιήσει ήδη υπάρχουσα γνώση ή αν θα εξερευνήσει καινούριες καταστάσεις με στόχο τη μεγαλύτερη πιθανή ανταμοιβή. Η εξερεύνηση είναι πολύ σημαντική αφού με αυτό τον τρόπο ο πράκτορας ανακαλύπτει νέες καταστάσεις που πιθανόν να επιφέρουν μεγαλύτερες ανταμοιβές. Δεν είναι δυνατό

όμως να εξερευνά συνεχώς γι' αυτό θα πρέπει να χρησιμοποιεί και τις γνώσεις του. Η πιο συνηθισμένη μέθοδος για να το πετύχουμε αυτό και να εξισορροπηθούν η εξερεύνηση με την αξιοποίηση είναι η **e-greedy** στην οποία με μια πιθανότητα $0 \leq \epsilon \leq 1$ ο πράκτορας επιλέγει μια τυχαία ενέργεια στην κατάσταση που βρίσκεται και με πιθανότητα $1-\epsilon$ επιλέγει την ενέργειά του με βάση τις γνώσεις του και τον αλγόριθμο που χρησιμοποιεί.

2.6 Θεωρία Παιγνίων

Η γενική ιδέα της θεωρίας παιγνίων είναι η μοντελοποίηση μερικών πράξεων και στρατηγικών ως ένα παιχνίδι μεταξύ διαφόρων παικτών. Έτσι ένα παιχνίδι είναι μια μαθηματική συνάρτηση η οποία περιγράφει τις συνέπειες της κάθε πράξης των παικτών κατά τη διάρκεια της προσπάθειας τους να αποκομίσουν όσο περισσότερο κέρδος μπορούν. Ο κάθε παίκτης ακολουθεί την δική του στρατηγική στο παίγνιο αλλά το κέρδος του εξαρτάται και από τις επιλογές του/των αντιπάλου/ων.

Ο όρος «Θεωρία Παιγνίων» παραπέμπει σε επιτραπέζια παίγνια, όπως το σκάκι, το τάβλι κλπ., αυτό οφείλεται στο ότι πολλές οικονομικές, πολιτικές, στρατιωτικές και άλλες συγκρούσεις μπορούν να παρομοιαστούν με τις πράξεις αυτών των παιγνίων. Σε κάθε παίγνιο υπάρχουν οι διάφοροι παίκτες. Ο κάθε παίκτης έχει στην διάθεση του μερικές επιλογές τις οποίες χρησιμοποιεί με βάση την στρατηγική που ακολουθεί. Επακόλουθο των πράξεων του σε συνδυασμό με τις πράξεις και των υπόλοιπων παικτών διατυπώνονται ως συναρτήσεις απώλειας ή κέρδους.

Τα παίγνια ξεχωρίζουν μεταξύ τους με βάση τους κανόνες που τα διέπουν. Οι κανόνες αυτοί είναι γνωστοί σε όλους τους παίκτες του παιχνιδιού και με βάση αυτούς οι παίκτες γνωρίζουν τις δυνατότητες που έχουν. Σε κάθε κατάσταση του παίγνιου οι παίκτες έχουν μερικές προκαθορισμένες κινήσεις τις οποίες μπορούν να ακολουθήσουν από τις οποίες και διαλέγουν ποια θα ακολουθήσουν σε κάθε γύρο.

Ανάλογα με τις συναρτήσεις αμοιβών των παικτών μπορεί να γίνει μια κατηγοριοποίηση των παιχνιδιών. Όταν όλα τα rewards των παικτών αθροίζονται στο μηδέν τότε το παίγνιο καλείται μηδενικού αθροίσματος (zero-sum game). Στα παίγνια αυτά η νίκη συγκεκριμένων παικτών συνεπάγεται την ήττα των παικτών με αντίθετους στόχους.

Αντίθετα, υπάρχουν και τα παίγνια γενικού αθροίσματος στα οποία οι παίκτες με τις κατάλληλες επιλογές έχουν τη δυνατότητα να τερματίσουν όλοι κερδισμένοι, δηλαδή, η νίκη κάποιων παικτών μπορεί να συνυπάρξει με αυτή των παικτών με διαφορετικές/αντίθετες προθέσεις.

Ακόμη, τα παίγνια μπορούν να ταξινομηθούν με βάση τον αριθμό των παικτών που συμμετέχουν σε αυτά. Για να υπάρξει ανταγωνισμός απαιτούνται τουλάχιστον 2 παίκτες και αυτά ονομάζονται παίγνια δύο παικτών. Υπάρχουν επίσης παίγνια με περισσότερους από δύο παίκτες τα οποία ονομάζονται παίγνια n παικτών στα οποία οι παίκτες μπορούν να δημιουργήσουν ομάδες ή/και συμμαχίες.

Επιπρόσθετα, έχουμε τα στατικά και τα δυναμικά παίγνια, στα πρώτα η σειρά με την οποία οι παίκτες λαμβάνουν τις αποφάσεις δεν παίζει ρόλο και στα δεύτερα η σειρά έχει σημασία.

Τα παίγνια αρχικά υπήρξαν ως παίγνια «κανονικής μορφής» (normal form) τα οποία και εκτελούνταν μόνο μια φορά και οι παίκτες μπορούσαν να βρίσκονται μόνο σε μία κατάσταση. Έπειτα όμως, υπήρξε μια ανάγκη για επαναλαμβανόμενα παίγνια (repeated games), στα οποία η μακροχρόνια στρατηγική ήταν αυτή που οδηγούσε τους παίκτες σε καλύτερες ισορροπίες ανεξάρτητα από τα αποτελέσματα των επιλογών/ συνεπειών των αποφάσεων της κάθε κατάστασης. Αυτές οι ισορροπίες ονομάζονται ισορροπίες Nash (Nash, 1950), και ουσιαστικά αποτελούν στρατηγικές οι οποίες είναι ωφέλιμες τόσο και για το προσωπικό κέρδος του παίκτη όσο και για των αντιπάλων του.

2.7 Bargaining Game

Το bargaining game Rubinstein (1982) αποτελεί ένα παίγνιο γενικού αθροίσματος, δηλαδή μπορούν να υπάρξουν και οι 2 παίκτες νικητές/κερδισμένοι. Στο συγκεκριμένο παίγνιο όσο πιο γρήγορα οι 2 παίκτες συνεργαστούν τόσο πιο μεγάλο μπορεί να είναι το κέρδος και για τους 2. Αν δεν συνεργάζονται καθόλου οι παίκτες τότε θα καταλήξουν με μηδενικό κέρδος και οι 2.

Το πιο συνηθισμένο παράδειγμα του παιχνιδιού είναι το εξής:



Σχήμα 2.6.1

Σχήμα αλληλεπίδρασης δύο παικτών στο Bargaining Game

Αρχικά υπάρχει το καλάθι μας το οποίο και περιέχει ολόκληρο το ποσό για το οποίο γίνεται η διαπραγμάτευση. Ο παίκτης Α κάνει μια προσφορά στον παίκτη Β και αναλόγως αυτός την αποδέχεται ή κάνει μια αντιπροσφορά. Με την απάντηση του παίκτη Β ο γύρος τελειώνει, αν η απάντηση ήταν θετική τότε ο κάθε παίκτης παίρνει το ανάλογο ποσό, αν ήταν αρνητική ο παίκτης Α καλείται να απαντήσει στην αντιπροσφορά και το ποσό μειώνεται.

Έχουμε 2 παίκτες, τον παίκτη Α και τον παίκτη Β. Σκοπός τους είναι να μοιράσουν το ποσό X . Αρχικά το ποσό X έχει την αρχική του αξία, με βάση την οποία ο παίκτης Α κάνει την προσφορά του (την προσφορά α) στον παίκτη Β. Ο παίκτης Β αποφασίζει εάν θα αποδεχτεί την προσφορά, δηλαδή αυτός να πάρει α μονάδες από το ποσό X και ο παίκτης Α να πάρει τις $X-\alpha$ μονάδες, ή επιλέγει να απορρίψει την προσφορά του Α και να κάνει μια αντιπροσφορά β . Το παιχνίδι όμως μετά από κάθε γύρο αλλάζει το ποσό, αν δεν αποδέχτηκε ο παίκτης και έκανε μια αντιπροσφορά, αφαιρείται μια μονάδα από το ποσό X . Οπότε στον επόμενο γύρο το ποσό X πλέον θα είναι $X-1$ για το οποίο και οι 2 παίκτες πρέπει να συμφωνήσουν σε μια ικανοποιητική προσφορά και για τους 2. Αν οι 2 παίκτες κρατούν μια αδιάλλακτη στάση και συνεχώς αρνούνται την προσφορά του αντιπάλου τότε το ποσό X θα εξαντληθεί και δεν θα πάρει κανείς τίποτα.

Είναι προφανές πως όσο πιο γρήγορα γίνει αποδοχή μιας προσφοράς από τον 1 παίκτη τότε το συνολικό κέρδος και για τους 2 θα είναι μεγαλύτερο. Το ιδανικό σενάριο είναι από την πρώτη προσφορά που θα γίνει ο διαμοιρασμός να είναι δίκαιος και κατά πάσα

πιθανότητα στην μέση, δηλαδή 50-50 και ο αντίπαλος να δεχτεί την προσφορά χωρίς να έχει αφαιρεθεί καμία μονάδα από το ποσό.

Στο παιχνίδι αυτό προστίθεται και η έννοια της επανάληψης με την επαναφορά του ποσού στον αρχικό του αριθμό και την επανάληψη του παιχνιδιού. Δηλαδή, όταν το ποσό μηδενιστεί ή ένας παίκτης αποδεκτεί την προσφορά του άλλου παίκτη, παίρνουν και οι δύο το ποσό/ την επιβράβευση τους και το pot επαναγεμίζει. Με αυτό τον τρόπο οι παίκτες μπορούν να χρησιμοποιήσουν τις εμπειρίες τους από προηγούμενες επαναλήψεις του παιχνιδιού για να βρουν τι τους συμφέρει περισσότερο και να βελτιώσουν τις προσφορές τους για το κοινό καλό.

Επιλέξαμε αυτό το παίγνιο για να χρησιμοποιηθεί ως το παίγνιο στο οποίο θα μαθαίνουν οι 2 πράκτορές μας διότι αντικατοπτρίζει σε μεγάλο βαθμό την πραγματικότητα. Αφού με την πάροδο του χρόνου και την αδιάλλακτη στάση των 2 πλευρών, δηλαδή την άρνηση των προσφορών της απέναντι πλευράς ή ακόμη και την προσφορά πολύ χαμηλών ποσοστών σε σχέση με το επιθυμητό του αντίπαλου οι 2 πλευρές συνεχίζουν να τιμωρούνται με τις συνέπειες της ισχύουσας κατάστασης.

Τέλος, αυτό το παίγνιο προσφέρεται αφού έχουμε ένα σταθερό αρχικό pot το οποίο οι 2 πλευρές προσπαθούν να διαμοιράσουν και συνεχώς, κατά τις συνομιλίες, η κάθε πλευρά είναι πιθανόν να κάνει μια προσφορά την οποία η άλλη πλευρά καλείται να αποδεκτεί ή να απορρίψει ανάλογα.

2.8 Κυπριακό Πρόβλημα

Το Κυπριακό πρόβλημα, ή απλώς το «Κυπριακό» όπως αποκαλείται αφού μετά από δεκαετίες αποτελεί το κύριο πρόβλημα αυτού του τόπου, είναι ένα πρόβλημα διεθνούς δικαίου το οποίο άρχισε να εμφανίζεται πιο έντονα από το 1964 με τις πρώτες διακοινοτικές ταραχές στην Κύπρο και συνεχίζεται μέχρι και σήμερα, το 2022.

Αποτελεί ένα πολύ σημαντικό αλλά συνάμα και περίπλοκο πρόβλημα. Υπάρχουν διάφορες πτυχές του θέματος για τις οποίες γίνονται συζητήσεις όπως είναι το εδαφικό η διακυβέρνηση κ.α. Επιπρόσθετα, το πρόβλημα αυτό εμπλέκει τις 2 πλευρές ΕΚ και ΤΚ

αλλά όχι μόνο. Εμπλέκονται και άλλες χώρες οι οποίες έχουν τα δικά τους συμφέροντα με βάση την έκβαση των διαπραγματεύσεων. Οι δύο εξωτερικοί παράγοντες που έχουν εξ ολοκλήρου συμφέρον από την μια πλευρά ΕΚ και ΤΚ είναι η Ελλάδα και η Τουρκία αντίστοιχα. Εμπλέκονται όμως και η Αγγλία, ως εγγυήτρια δύναμη, η ΗΠΑ και τα Ηνωμένα Έθνη ως εξωτερικές δυνάμεις και οργανισμοί.

Η εδαφική πτυχή του προβλήματος αρχίζει κυρίως από το 1974 όπου μετά την τουρκική στρατιωτική εισβολή στην Κύπρο αφού από τότε η Τουρκία κρατεί στην κατοχή της περίπου το 37% του Κυπριακού εδάφους. Πιο συγκεκριμένα, η Τουρκία ελέγχει το Βόρειο κομμάτι της Κυπριακής Δημοκρατίας όπου κατοικούν κυρίως ΤΚ οι οποίοι μεταφέρθηκαν στις περιοχές αυτές μετά την εισβολή μαζί με Τούρκους έποικους που ήρθαν από την Τουρκία. Στο υπόλοιπο νησί, εξαιρουμένου του περίπου 3% που κατέχουν οι Αγγλικές βάσεις και ένα ποσοστό εδάφους το οποίο δόθηκε από την ΚΔ στον ΟΗΕ για να γίνει νεκρά ζώνη (πράσινη γραμμή), κατοικούν ΕΚ. Η τουρκική πλευρά αποκαλεί αυτά τα εδάφη ξεχωριστό κράτος, το οποίο όμως δεν αναγνωρίζεται από τη διεθνή κοινότητα ή κάποιο άλλο κράτος εκτός της Τουρκίας.

Οι δύο πλευρές επιδιώκουν από τότε να βρεθεί μια λύση στο πρόβλημα με την ΕΚ πλευρά να διεκδικεί την εδαφική ακεραιότητα της Κυπριακής Δημοκρατίας και την ΤΚ πλευρά να επιδιώκει ποσοστά διακυβέρνησης ή και εδάφη που να ανήκουν στην Τουρκία. Οι πιο πρόσφατες συνομιλίες στις οποίες υπήρξε μια πιο ξεκάθαρη εικόνα για το εδαφικό κομμάτι είναι αυτές που έγιναν στην Ελβετία το 2017. Σε αυτή την προσπάθεια εξεύρεσης λύσης οι 2 πλευρές κατέθεσαν με αριθμούς για πρώτη φορά τα ποσοστά με τα οποία είναι ευχαριστημένοι. Η ΤΚ πλευρά επιθυμεί να κρατήσει το 29.2% της Κύπρου καθώς η ΕΚ πλευρά διεκδικεί το 8% από το 36.2% που κατέχει τώρα η ΤΚ πλευρά. Η διαφορά δηλαδή στα ποσοστά που επιθυμούν ανέρχεται στο 1%. Δηλαδή, αν η ΕΚ πάρει το ποσοστό που επιθυμεί, η ΤΚ πλευρά μένει με 28.2%, 1% λιγότερο από το επιθυμητό για την ΤΚ πλευρά. Αν τώρα, η ΤΚ πλευρά πάρει το ποσοστό που επιθυμεί, η ΕΚ πλευρά μένει με ποσοστό 7%, 1% λιγότερο από το επιθυμητό για την ΕΚ πλευρά.

Τα ποσοστά πληθυσμού στο νησί σε απογραφή που είχε γίνει το 1960, μερικά χρόνια πριν την εισβολή δηλαδή ήταν 77.1% ΕΚ, 18.2% ΤΚ και το υπόλοιπο ποσοστό άλλοι. Ο λόγος που γίνεται αναφορά σε αυτή την απογραφή είναι διότι μετά την εισβολή η

Τουρκία άρχισε μια συνεχή προσπάθεια μεταφοράς Τούρκων έποικων στα κατεχόμενα εδάφη. Είναι ξεκάθαρο πως τα ποσοστά πληθυσμού είναι δυσανάλογα με τα ποσοστά εδάφους που διεκδικεί η ΤΚ πλευρά.

Όπως έχει ήδη αναφερθεί, ο λόγος για τον οποίο θα ασχοληθούμε με την εδαφική πτυχή του προβλήματος είναι διότι το εδαφικό προσφέρεται λόγω του ότι πλέον έχουμε συγκεκριμένους αριθμούς τους οποίους διεκδικεί η κάθε πλευρά. Είναι ξεκάθαρο δηλαδή το ποιες είναι οι προσδοκίες της κάθε πλευράς κάτι το οποίο δεν ισχύει και για την διακυβέρνηση.

2.9 Προηγούμενη εργασία

Αρκετή δουλειά έχει γίνει τα τελευταία χρόνια για μοντελοποίηση διαφορετικών προβλημάτων με τη χρήση Ενισχυτικής Μάθησης Πολλαπλών Πρακτόρων.

Αρχικά, ο Βασιλειάδης (2007) ερεύνησε εάν μπορεί να υπάρξει συνεργασία μεταξύ δύο πρακτόρων οι οποίοι εκπαιδεύτηκαν με Ενισχυτική μάθηση μέσα σε ένα παίγνιο γενικού αθροίσματος.

Ο Καραολής (2010) εργάστηκε στη χρήση παίγνιου γενικού αθροίσματος με τη χρήση ΕΜΠΠ για μια προσέγγιση του προβλήματος του Κυπριακού. Ο Καραολής ασχολήθηκε με ολόκληρο το πρόβλημα του Κυπριακού και όχι αποκλειστικά με μια πτυχή του. Επίσης, η αντιστοίχιση έγινε για διαφορετικές περιόδους. Αποτέλεσμα αυτού ήταν για την κάθε περίοδο να υπάρχουν πολλοί παράγοντες που έπρεπε να ληφθούν υπόψη και να αντιστοιχηθούν σε ένα και μόνο ενισχυτικό σήμα. Επομένως, τα αποτελέσματα ήταν πολύ γενικά και δεν είναι εύκολο να εξαχθεί συγκεκριμένο συμπέρασμα για μια πτυχή του προβλήματος. Το παίγνιο το οποίο χρησιμοποίησε ήταν το επαναλαμβανόμενο δίλημμα του φυλακισμένου. Στο ίδιο θέμα με τη χρήση του ίδιου παιγνίου είχε γίνει δουλειά και από τον Lumsden (1973).

Ακόμη, οι Nowé et al. (2012) ανέλυσαν τις εφαρμογές της ενισχυτικής μάθησης σε συστήματα πολλαπλών πρακτόρων, εξήγησαν βασικές μεθόδους εκμάθησης με τη χρήση της θεωρίας παιγνίων και παρουσίασαν την πολυπλοκότητα που προκύπτει από τέτοιου είδους συστήματα.

Επιπρόσθετα, ο Ιωάννου (2017) έκανε μια προσέγγιση του κυπριακού και πιο συγκεκριμένα του εδαφικού προβλήματος με τη χρήση του Bargaining game. Χρησιμοποίησε απλουστευμένα τα ποσοστά των δύο πλευρών και χωρίς την προσπάθεια εισαγωγής των εξωτερικών παραγόντων που έχουν συμφέροντα μέσω του κυπριακού προβλήματος. Ο τρόπος υλοποίησης του συστήματος και της αντιστοίχισης του Κυπριακού στο σύστημα ήταν πολύ χονδρικός. Ο τρόπος που έγινε η αντιστοίχιση ήταν αρκετά διαφορετικός από αυτόν που θεωρήσαμε εμείς ως πιο ρεαλιστικό. Υπήρχαν μόνο 37 καταστάσεις για τους 2 παίκτες στις οποίες διαπραγματεύονταν για 1% γης. Συμβιβασμός μπορούσε να υπάρξει μόνο στον πρώτο γύρο έτσι το αποτέλεσμα των εκπαιδεύσεων ήταν οι πράκτορες να μαθαίνουν να προτείνουν ένα συγκεκριμένο ποσοστό (αυτό που θα έφερνε τον συμβιβασμό) και να αποδέχονται την πρώτη προσφορά, με σκοπό να καταλήξουν σε συμβιβασμό.

Κεφάλαιο 3

Σχεδιασμός και Υλοποίηση

3.1 Εισαγωγή	27
3.2 Σχεδιασμός Συστήματος	29
3.3 Ενσωμάτωση Άδηλων Πηγών Ενίσχυσης	30
3.4 Υλοποίηση	32

3.1 Εισαγωγή

Με σκοπό την εξερεύνηση των πιο πάνω, δημιουργήσαμε ένα σύστημα χρησιμοποιώντας την γλώσσα προγραμματισμού JAVA. Η επιλογή της γλώσσας έγινε λόγω του ότι, η συγκεκριμένη, μας προσφέρει αντικειμενοστρέφεια, και παρέχει τις δυνατότητες που απαιτούνταν και ικανοποιητική επίδοση. Το σύστημα αυτό έχει την δυνατότητα να λειτουργεί με τους απαιτούμενους πράκτορες για το πρόβλημά μας, χρησιμοποιώντας EM για την εκπαίδευσή τους μέσω και της χρήσης του Bargaining Game.

Αρχικά, οι 2 πράκτορές μας είναι οι EK και οι TK οι οποίοι επιδιώκουν να αποκομίσουν το επιθυμητό, για τους ίδιους, αποτέλεσμα (ποσοστό γης). Οι αλγόριθμοι που χρησιμοποιούν οι πράκτορες για την εκπαίδευσή τους είναι ο Q-Learning και ο SARSA με τη χρήση e-greedy τακτικής. Οι δύο αυτοί αλγόριθμοι επεκτείνουν την κλάση Agent που αποτελεί την γενική abstract κλάση και περιέχει τις απαραίτητες μεθόδους για να λειτουργήσει ο κάθε πράκτορας. Υπάρχει η δυνατότητα επέκτασης της κλάσης Agent με σκοπό τη χρήση οποιουδήποτε άλλου αλγορίθμου.

Αρχικός στόχος ήταν η μοντελλοποίηση του συστήματος και η εξερεύνηση της επιρροής της κάθε παραμέτρου στην πιθανότητα οι 2 πράκτορες να οδηγηθούν στον συμβιβασμό

(compromise), στον μη-συμβιβασμό (no compromise) ή στη νίκη του ενός παίκτη και την ήττα του άλλου (win-EK, win-TK). Οι πράκτορες μέσα από πολλές επαναλήψεις και προσπάθειες προσπαθούν να εντοπίσουν τις επιλογές που τους επιφέρουν την περισσότερη επιβράβευση.

Έπειτα, σκοπός ήταν να εντάξουμε σε αυτό το πρόγραμμα την έννοια της άδηλης πηγής ενίσχυσης από την ψυχολογία (Γεωργίου, 2007; Deci, 1972; Ryan & Deci, 2000) έτσι ώστε τα αποτελέσματα των πρακτόρων να είναι πιο κοντά στην πραγματικότητα. Βέβαια, θέλουμε επίσης να εξερευνήσουμε κατά πόσο αλλά και με ποιους τρόπους (επιβράβευση ή τιμωρία σε συγκεκριμένο παίκτη) οι εξωτερικοί παράγοντες έχουν τη δυνατότητα να επηρεάσουν το αποτέλεσμα των 2 παικτών, κυρίως ως προς τον συμβιβασμό.

Οι 5 έμμεσα εμπλεκόμενοι που χρησιμοποιήσαμε (Ελλάδα, Τουρκία, Η.Π.Α, Ο.Η.Ε., Η.Β.) δεν αποτελούν άμεσους παίκτες στο παιχνίδι αλλά λειτουργούν ως πηγές ενίσχυσης ή τιμωρίας των 2 βασικών παικτών (EK,TK) όπως ουσιαστικά συμβαίνει και στην πραγματικότητα. Ο κάθε ένας από τους 5 έμμεσα εμπλεκόμενους ενισχύει ή τιμωρεί διαφορετικά την κάθε πράξη των 2 παικτών και αυτό εξαρτάται κυρίως από τα δικά τους συμφέροντα και την εξέλιξη την οποία επιδιώκουν για το πρόβλημα. Ο κάθε πράκτορας μπορεί να έχει το δικό του βάρος για τους εξωτερικούς παράγοντες, δηλαδή επιλέγει το κατά πόσο θα τον επηρεάζουν οι επιβραβεύσεις ή οι τιμωρίες από τους εξωτερικούς παράγοντες. Την παράμετρο αυτή την ορίζουμε ως το βάρος που δίνει ένας πράκτορας στο ενισχυτικό σήμα που επιστρέφουν οι εξωτερικοί παράγοντες. Αυτό σημαίνει πως ένα βάρος (external factors weight) με τιμή μηδέν (0) ισοδυναμεί με την εξ ολοκλήρου απουσία των εξωτερικών παραγόντων, αντιθέτως, ένα βάρος ίσο με 1 ισοδυναμεί με πράκτορα ο οποίος λαμβάνει στο μέγιστο υπόψιν του τα ενισχυτικά σήματα των εξωτερικών παραγόντων.

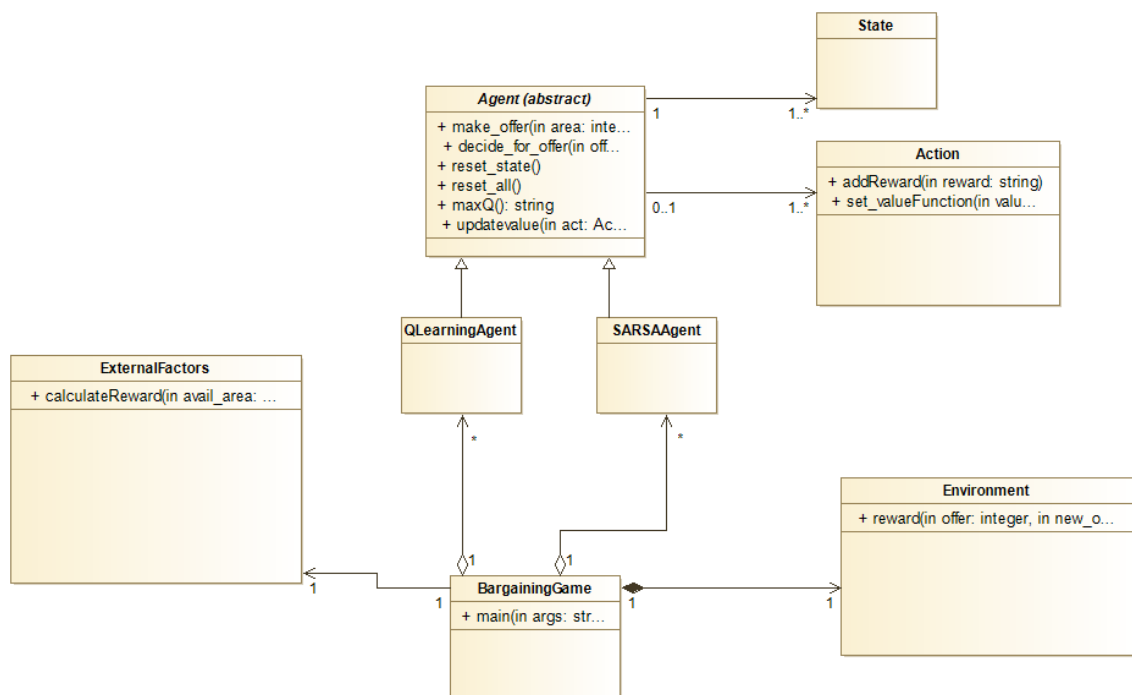
Επιπρόσθετα, υλοποιήθηκε η δυνατότητα επιπρόσθετης αμοιβής από τους εξωτερικούς παράγοντες για συγκεκριμένα αποτελέσματα των 2 παικτών. Οι αμοιβές αυτές αφορούν όλες τις διαδοχικές επιλογές/αποφάσεις τις οποίες έκανε ο/οι παίκτης/ες για να καταλήξει στο συγκεκριμένο αποτέλεσμα και όχι μόνο την τελευταία του επιλογή. Αυτή η επιλογή αποτελεί κυρίως πειραματική περιέργεια και υπάρχει η επιλογή της απενεργοποίησής

της χωρίς να επηρεάζει τις εξωτερικές πηγές ενίσχυσης. Τέλος, όλες οι παράμετροι που απαιτούνται για τις διάφορες επιλογές του συστήματος που αναφέρθηκαν πιο πάνω δίνονται μέσω ενός αρχείου (parameters.txt).

3.2 Σχεδιασμός Συστήματος

Όπως έχει προαναφερθεί επιλέξαμε τη γλώσσα προγραμματισμού JAVA κυρίως διότι προσφέρει αντικειμενοστρέφεια. Δημιουργούμε αντικείμενα του κατάλληλου Agent που επιθυμούμε και αυτός δρα μέσα στο παιχνίδι (bargaining) αλληλοεπιδρώντας με το περιβάλλον.

Στο σχήμα 3.2.1 φαίνονται οι διάφορες κλάσεις του συστήματος και το πώς αυτές συνδέονται μεταξύ τους.



Σχήμα 3.2.1

Σχήμα Διάγραμμα κλάσεων

Στο σχήμα αυτό παρουσιάζονται οι κλάσεις του συστήματος, κάποιες βασικές συναρτήσεις και η σχέσεις που έχουν οι κλάσεις μεταξύ τους.

Αρχικά, έχουμε τις κλάσεις QLearningAgent και SARSAAgent οι οποίες κληρονομούν την αφαιρετική κλάση Agent και υλοποιούν όλες τις απαραίτητες συναρτήσεις. Οι δύο αυτές κλάσεις είναι οι αλγόριθμοι που χρησιμοποιούμε, ωστόσο μπορεί να υλοποιηθεί οποιοσδήποτε άλλος αλγόριθμος. Το μόνο που χρειάζεται είναι να υλοποιεί όλες τις συναρτήσεις της abstract κλάσης (κάνοντάς την extend αυτό είναι απαραίτητο).

Οι agent αυτοί δημιουργούνται ανάλογα με τις επιλογές που δίνουμε μέσω του parameters.txt file. Μετά την ανάγνωση των παραμέτρων από την κλάση Bargaining game αν επιλέξουμε να έχουμε Q-learning agent, η BargainingGame κλάση, η οποία ουσιαστικά χειρίζεται όλο το παιχνίδι, θα δημιουργήσει δύο αντικείμενα της κλάσης QLearningAgent, ένα για τον κάθε πράκτορα (EK και TK). Έπειτα, αρχίζει η διαδικασία του παίγνιου.

Ο κάθε agent έχει τον δικό του πίνακα από όλα τα states στα οποία μπορεί να βρεθεί. Το κάθε ένα από αυτά τα states (καταστάσεις) περιέχει τα actions τα οποία ο agent έχει τη δυνατότητα να εκτελέσει στη συγκεκριμένη κατάσταση.

Τέλος, η κλάση environment είναι η κλάση που είναι υπεύθυνη για να τιμωρήσει ή να επιβραβεύσει τους agents ανάλογα με τις αποφάσεις που παίρνουν. Η κύρια συνάρτηση της κλάσης αυτής δέχεται ως είσοδο την προσφορά, την πιθανή αντιπροσφορά, το διαθέσιμο προς συζήτηση έδαφος και τον πράκτορα τον οποίο αποφασίζει έτσι ώστε να υπολογίσει το ενισχυτικό σήμα το οποίο θα επιστρέψει.

3.3 Ενσωμάτωση Άδηλων Πηγών Ενίσχυσης

Επιπρόσθετα, από την προσομοίωση του προβλήματος, όπως περιεγράφηκε στο κεφάλαιο 3.2, σκοπός της διπλωματικής ήταν και η διερεύνηση της συμβολής που έχουν οι εξωτερικοί παράγοντες στην έκβαση των διαπραγματεύσεων. Οι παράγοντες αυτοί, προστέθηκαν ως άδηλες πηγές ενίσχυσης οι οποίες επιστρέφουν το δικό τους ενισχυτικό σήμα στους πράκτορες αναλόγως του προσωπικού τους συμφέροντος και του κατά πόσο η απόφαση του agent είναι συμβατή με τα δικά τους συμφέροντα. Ο agent, με τη χρήση

μιας παραμέτρου που δίνεται στο parameters.txt επιλέγει κατά πόσο θα λαμβάνει υπόψιν του τα ενισχυτικά αυτά σήματα από τους εξωτερικούς παράγοντες.

Υπάρχουν οι χώρες οι οποίες επιθυμούν το κέρδος προς την δική τους πλευρά, δηλαδή η Τουρκία για τους ΤΚ και η Ελλάδα για τους ΕΚ. Πέρα από αυτές όμως έχουμε επίσης χώρες και οργανισμούς όπως οι Η.Π.Α, το Η.Β. και ο Ο.Η.Ε. οι οποίοι επιθυμούν να υπάρξει ο μέγιστος δυνατός συμβιβασμός, παράλληλα φυσικά με την εξυπηρέτηση κάποιων δικών τους συμφερόντων.

Για να επιτευχθεί αυτός ο στόχος δημιουργήσαμε την κλάση ExternalFactors. Η δημόσια συνάρτηση αυτής της κλάσης δέχεται ως είσοδο το διαθέσιμο προς διαπραγμάτευση έδαφος, τον πράκτορα τον οποίο κάνει την προσφορά, τον πράκτορα που αποφασίζει για την προσφορά, το ποσό της προσφοράς αλλά και την απόφαση του πράκτορα.

Εσωτερικά η κλάση αυτή χρησιμοποιώντας τις τιμές εισόδου υπολογίζει το ενισχυτικό σήμα του κάθε εξωτερικού παράγοντα και υπολογίζει το συνολικό σήμα από όλους το οποίο και επιστρέφει στον πράκτορα.

Τέλος, υπάρχει η δυνατότητα επιπρόσθετης αμοιβής από τους εξωτερικούς παράγοντες για συγκεκριμένα αποτελέσματα των 2 παικτών. Πιο συγκεκριμένα μπορεί να ενεργοποιηθεί η επιλογή με την οποία οι πράκτορες επιβραβεύονται για όλες τις διαδοχικές ενέργειες που ακολούθησαν και τους οδήγησαν σε συμβιβασμό και να τιμωρηθούν για την σειρά ενεργειών που τους οδήγησαν σε μη συμβιβασμό. Ακόμη, η άλλη επιλογή ενεργοποιεί την μονόπλευρη τιμωρία. Δηλαδή, αν ο ένας πράκτορας κατέληξε σε νίκη με αυτή την επιλογή ενεργοποιημένη θα τιμωρηθούν όλες οι προηγούμενες επιλογές του. Όλα αυτά προαπαιτούν την ύπαρξη των εξωτερικών παραγόντων (`external_weight>0`).

3.4 Υλοποίηση

Προτού αναλύσουμε τον κώδικα μέσω ψευδοκώδικα ας δούμε ποιες είναι οι καταστάσεις (state) στις οποίες μπορούν να βρεθούν οι 2 πράκτορές μας και τις ενέργειες στις οποίες μπορούν να προβούν.

Όπως έχει αναφερθεί το ποσοστό εδάφους για το οποίο γίνεται η διαπραγμάτευση αποτελεί το 36.2% του νησιού. Ορίσαμε την ποσότητα που θα μειώνεται σε κάθε γύρο που περνάει στο παιχνίδι ως 0.1%. Για ευκολία θέσαμε το αρχικό ποσό συζήτησης σε 362 μονάδες (κάθε μονάδα αντιστοιχεί σε 0.1% γης) και σε κάθε γύρο αυτό το ποσό μειώνεται κατά 1. Επακόλουθο του πιο πάνω είναι ο κάθε agent να μπορεί να βρεθεί σε 362 διαφορετικές καταστάσεις, μία κατάσταση για κάθε μονάδα συζήτησης.

Η κατάσταση ορίζεται από τις διαθέσιμες μονάδες εδάφους που υπάρχουν, δηλαδή το Pot για το οποίο διαπραγματεύονται οι 2 πράκτορες. Στην κάθε κατάσταση ο πράκτορας έχει 2 βασικές ενέργειες που μπορεί να εκτελέσει, είτε να αποδεχθεί είτε να απορρίψει μια προσφορά. Επιπρόσθετα όμως, χωρίζουμε τις 2 αυτές ενέργειες σε 2 υποκατηγορίες:

1. Αν η προσφορά είναι μικρότερη από το επιθυμητό ποσοστό του πράκτορα, τότε έχει τις δύο αυτές επιλογές.
2. Αν η προσφορά είναι ίση ή μεγαλύτερη από το επιθυμητό ποσοστό τότε και πάλι έχει αυτές τις δύο επιλογές.

Αυτή η διαφοροποίηση έχει ως αποτέλεσμα ο πράκτορας να έχει 4 πιθανές ενέργειες σε κάθε κατάσταση όμως, οι 2 είναι αμοιβαία αποκλειόμενες από τις άλλες 2 και έτσι καταλήγει με 2 ενέργειες διαθέσιμες σε κάθε κατάσταση.

State	Action/Decision		Value function
Available Area/Pot	Offer		
362	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0
	>= satisfied	Accept	0

361		Refuse	0
	< satisfied	Accept Refuse	0 0
360	>= satisfied	Accept Refuse	0 0
	< satisfied	Accept Refuse	0 0
359	>= satisfied	Accept Refuse	0 0
	< satisfied	Accept Refuse	0 0
.	.	.	.
.	.	.	.
.	.	.	.
2	>= satisfied	Accept Refuse	0 0
	< satisfied	Accept Refuse	0 0
1	>= satisfied	Accept Refuse	0 0
	< satisfied	Accept Refuse	0 0
0	>= satisfied	Accept Refuse	0 0
	< satisfied	Accept Refuse	0 0

Πίνακας 3.4.1

Πίνακας Καταστάσεων και ενεργειών

Στον πίνακα αυτό απεικονίζονται οι καταστάσεις που μπορεί να βρεθεί ένας παίκτης, από την 362^η κατάσταση μέχρι και την κατάσταση 0. Σε κάθε κατάσταση βλέπουμε τις επιλογές ενεργειών που έχει ο πράκτορας και την τιμή της συνάρτησης αξίας.

Ο πίνακας 3.4.1 παρουσιάζει το σύνολο των καταστάσεων και τις ενέργειες στις οποίες μπορεί να βρεθεί και να εκτελέσει, αντίστοιχα, ένας από τους πράκτορες. Βλέπουμε πως σε κάθε κατάσταση ο πράκτορας έχει να επιλέξει μεταξύ 2 ενεργειών. Επίσης, η συνάρτηση αξίας έχει αρχικοποιηθεί και για τους 2 πράκτορες για όλες τις καταστάσεις και όλες τις ενέργειες στο 0.

Συνεχίζοντας με την λειτουργία του προγράμματος, σε πρώτο στάδιο έχουμε τον βρόχο για τις εποχές που θα τρέξει ο αλγόριθμός μας. Μέσα σε αυτό τον βρόχο γίνεται η μάθηση των πρακτόρων μας. Αρχικά, επιλέγεται τυχαία ποιος πράκτορας θα αρχίσει και στην συνέχεια αυτός κάνει την πρώτη του προσφορά στον άλλο πράκτορα. Ο πράκτορας ο οποίος δέχεται την προσφορά όπως φαίνεται και στον πίνακα 3.4.1 έχει τις επιλογές να αποδεχθεί ή να απορρίψει την προσφορά. Αν τώρα ο πράκτορας αποδεχθεί την προσφορά τότε η εποχή αυτή τελειώνει και αξιολογεί κατά πόσο υπήρξε νικητής, συμβιβασμός ή κανένα από αυτά. Στην αντίθετη περίπτωση, τώρα που ο πράκτορας απέρριψε την προσφορά, είναι αναγκασμένος να δώσει μία αντιπροσφορά. Αμέσως μετά την αντιπροσφορά του πράκτορα ο γύρος τελειώνει, αφαιρείται μια μονάδα εδάφους από το διαθέσιμο ποσό και προχωράμε στον επόμενο γύρο. Τώρα ο πράκτορας ο οποίος στον προηγούμενο γύρο είχε κάνει την προσφορά καλείται να απαντήσει στην αντιπροσφορά που έκανε ο αντίπαλός του. Η διαπραγμάτευση συνεχίζεται μέχρι να αποδεχθεί ένας από τους 2 μια προσφορά ή μέχρι να εξαντληθεί τελείως το διαθέσιμο ποσό εδάφους

Μετά την κάθε απόφαση ενός πράκτορα η πράξη του αυτή αξιολογείται. Το περιβάλλον με βάση τα δεδομένα της συγκεκριμένης κατάστασης, της προσφοράς και της απόφασης που πήρε ο πράκτορας επιστρέφει πίσω την τιμή του reward/penalty που θα χρησιμοποιηθεί για να ενημερωθεί η συνάρτηση αξίας του πράκτορα. Επιπλέον, αν το βάρος για τις ΑΠΕ δεν είναι μηδέν τότε θα χρησιμοποιηθεί και το ενισχυτικό σήμα που θα επιστρέψει η κλάση externalFactors για να ληφθεί και αυτό υπόψη στην ενημέρωση της συνάρτησης αξίας.

Ακόμη, αν οι μεταβλητές για τις επιπρόσθετες αμοιβές (όπως έχουν εξηγηθεί στο 3.3) είναι ενεργοποιημένες με το τέλος του γύρου, θα τιμωρηθούν ή θα επιβραβευθούν όλες οι πράξεις των 2 πρακτόρων ανάλογα με το αποτέλεσμα.

Όλα αυτά που αναλύονται πιο πάνω περιγραφικά φαίνονται στον πιο κάτω ψευδοκώδικα.

1. Ανάγνωση παραμέτρων
2. Δημιουργία πρακτόρων (Q-learning ή SARSA)
3. Αρχικοποίηση των μετρητών σε 0
4. Δημιουργία αντικειμένου περιβάλλοντος.
5. Δημιουργία 1 λίστας για κάθε πράκτορα όπου θα αποθηκεύεται το ιστορικό των πράξεων τους.
6. Επιλογή τυχαία του πράκτορα που θα αρχίσει πρώτος να κάνει τις προσφορές.
7. Για κάθε επανάληψη Iteration:
 - 7.1. Ενόσω διαθέσιμο έδαφος>0
 - 7.1.1. Αν όλο το ποσό είναι διαθέσιμο τότε κάνε προσφορά στον αντίπαλο
 - 7.1.2. Αλλιώς κάνε την αντιπροσφορά που αποφάσισες στον προηγούμενο γύρο.
 - 7.1.3. `Action = agentdeciding.decide_for_offer(offer,current_area)`
 - 7.1.4. Αν η απάντηση είναι αρνητική τότε αποφάσισε την αντιπροσφορά
 - 7.1.5. `reward= environment.rew(offer,decision,current_area,agent_deciding)`
 - 7.1.6. **Αν δεν θα υπάρξει χρήση των ΑΠΕ τότε το `external_weight` θα ισουται με 0 έτσι δεν θα λαμβάνει υπόψιν αυτό το `reward`: `reward+= external_weight *ExternalFactor.calculateReward (current_area,off,decision,agentDeciding,agentOffering)`**
 - 7.1.7. `action.setValuefunction(((1 - lr) * action.value_function) + agentdeciding.updatevalue(action, reward))`
 - 7.1.8. Αν η απόφαση ήταν αποδοχή υπολόγισε κατά πόσο αποτελεί και τερμάτισε την επανάληψη(πάμε στο επόμενο Iteration):
 - 7.1.8.1. `Compromise++`
 - 7.1.8.2. `Win EK++`
 - 7.1.8.3. `Win TK++`
 - 7.1.8.4. `No compromise++`
 - 7.1.9. `current_area--`

Ψευδοκώδικας 3.4.1

Ψευδοκώδικας συστήματος

Αρχικά γίνεται η αρχικοποίηση των μετρητών και των παραμέτρων. Στην συνέχεια επιλέγεται ο πράκτορας που θα αρχίσει τις προσφορές πρώτος και οι διαπραγματεύσεις ξεκινάνε. Ο πράκτορας αυτός κάνει μια προσφορά ή αντιπροσφορά στον αντίπαλο και ο αντίπαλος καλείται να αποφασίσει. Η απόφαση του αυτή μαζί με την τρέχουσα κατάσταση αξιολογείται από το περιβάλλον το οποίο και του επιστρέφει το ενισχυτικό του σήμα. Επιπρόσθετα, αν χρησιμοποιούνται οι ΑΠΕ λαμβάνοντας και αυτές τις παραμέτρους που χρειάζονται (κατάσταση, προσφορά, απόφαση) επιστρέφουν το δικό τους ενισχυτικό σήμα. Έπειτα, γίνεται η ενημέρωση της συνάρτησης αξίας. Τέλος, αν η απάντηση ήταν αποδοχή τότε η επανάληψη τερματίζεται και υπολογίζεται το τελικό αποτέλεσμα (συμβιβασμός, νίκη, μη συμβιβασμός), αν δεν ήταν αποδοχή, το διαθέσιμο ποσό μειώνεται κατά ένα και προχωράμε στον επόμενο γύρο.

Κεφάλαιο 4

Αντιστοίχιση Κυπριακού Προβλήματος στο Bargaining game

4.1 Εισαγωγή	37
4.2 Περιληπτικά τα κομμάτια του Κυπριακού Προβλήματος που μας ενδιαφέρουν	37
4.3 Αντιστοίχιση του Κυπριακού προβλήματος με το σύστημα	39
4.4 Αντιστοίχιση Εξωτερικών Παραγόντων με το σύστημα	43

4.1 Εισαγωγή

Όπως έχει ήδη αναφερθεί, στόχος της εργασίας αυτής ήταν η χρήση της ΕΜ και του bargaining game για την προσομοίωση ενός πραγματικού προβλήματος. Το «Κυπριακό Πρόβλημα» είναι ένα πρόβλημα το οποίο εμπλέκει αρκετές πτυχές, από τις οποίες εμείς επιλέξαμε να ασχοληθούμε με το εδαφικό κομμάτι. Στο κεφάλαιο αυτό θα αναλύσουμε το κομμάτι το οποίο μας ενδιαφέρει και στην συνέχεια θα εξηγήσουμε τον τρόπο με τον οποίο έγινε η αντιστοίχιση του στο σύστημα μας.

4.2 Περιληπτικά τα κομμάτια του Κυπριακού Προβλήματος που μας ενδιαφέρουν

Σε αυτό το πρόβλημα έχουμε τους 2 άμεσα εμπλεκόμενους, ΤΚ και ΕΚ, αλλά και έμμεσα εμπλεκόμενους από τους οποίους επιλέξαμε να ασχοληθούμε με τους 5 πιο κύριους, την Τουρκία, την Ελλάδα, το Η.Β., τις Η.Π.Α. και τον Ο.Η.Ε..

Οι έντονες διαμάχες μεταξύ των 2 πλευρών φαίνεται να αρχίζουν από τα τέλη του 1963 (Σακκάς, 2016), όταν μετά από κάποιες εισηγήσεις του Μακάριου για αναθεώρηση του Συντάγματος ξεκίνησαν διακοινοτικές διαταραχές στο νησί. Η Τουρκία, με πρόσχημα τις προθέσεις αυτές του Μακάριου, άρχισε να απειλεί για στρατιωτική εισβολή και

διχοτόμηση του νησιού. Η στρατηγική σημασία της Κύπρου ήταν τεράστια για τα αμερικανικά και κατά συνέπεια «ΝΑΤΟϊκά» συμφέροντα, όπως και για το ΗΒ που διατηρούσε στρατιωτικές βάσεις στο νησί. Οι ΗΠΑ επιθυμούσαν τη διατήρηση της κυριαρχίας τους στη θαλάσσια περιοχή της Μεσογείου. Οι εξελίξεις του 1964 φαίνονταν ως τροχοπέδη στα σχέδια αυτά και έτσι η Αμερική, μετά και την αποτυχία της διάσκεψης του Λονδίνου τον Ιανουάριο του 1964, θεωρούσε πως ο διαχωρισμός του νησιού σε δύο κοινότητες θα εξυπηρετούσε καλύτερα τα συμφέροντά της. Έπειτα, στο πρώτο και στο δεύτερο σχέδιο Άτσεσον, προτεινόταν η Ένωση της Κύπρου με την Ελλάδα και η παραχώρηση της Καρπασίας (περίπου 11%) στην Τουρκία ως Βάση, στο πρώτο σχέδιο κατά πλήρη κυριαρχία στο δεύτερο με μακροχρόνια εκμίσθωση (Σακκάς, 2016).

Προχωρώντας μερικά χρόνια μπροστά, η Τουρκική στρατιωτική εισβολή επιχειρείται τελικά στις 20 Ιουλίου το 1974. Η Τουρκία στήριξε την εισβολή αυτή στη Συνθήκη Εγγύησής του 1959 με το επιχείρημα της προστασίας της τουρκοκυπριακής κοινότητας απέναντι στην επικείμενη ένωση με την Ελλάδα. Η εισβολή ακολουθήθηκε από τη δεύτερη επιχείρηση του τουρκικού στρατού στις 14 Αυγούστου 1974 κατά την οποία η Τουρκία κατέλαβε το 36.2% του νησιού. Η Τουρκία προχώρησε στην ανακήρυξη κράτους στα κατεχόμενα εδάφη της ΚΔ, το 1983, το οποίο δεν αναγνωρίζεται από τους διεθνείς οργανισμούς και κανένα κράτος εκτός από την Τουρκία.

Μετά από χρόνια διαπραγματεύσεων και λόγω της πολυπλοκότητας του προβλήματος, μέχρι σήμερα, 48 χρόνια μετά (2022) δεν έχουν υπάρξει σημαντικές εξελίξεις που να έχουν εμφανή αποτελέσματα στο Κυπριακό.

Αφήνοντας πίσω πολλές συζητήσεις που έγιναν κατά τα μετα-εισβολής χρόνια φτάνουμε στο 2017, στη χρονιά κατά την οποία είχαμε τις συζητήσεις στη Ελβετία όπου και υπήρξαν για πρώτη φορά σημαντικές εξελίξεις όσο αφορά το εδαφικό. Στις συζητήσεις αυτές, και οι δύο πλευρές παρουσίασαν χάρτες στους οποίους υπήρχαν λεπτομερώς τα ποσοστά εδάφους που διεκδικούσε η κάθε πλευρά να τεθούν υπό την διοίκηση της. Αυτή ήταν μια πολύ μεγάλη εξέλιξη για τις συζητήσεις του Κυπριακού.

Τα ποσοστά που φαίνεται να απαιτούσαν οι 2 πλευρές ήταν αρκετά κοντά σε ομοφωνία, αφού η ΤΚ πλευρά ζητούσε το 29.2% του εδάφους ενώ η ΕΚ ζητούσε το 8% αφήνοντας

την TK με 28.2%. Η διαφορά, δηλαδή, μεταξύ των 2 επιθυμητών ποσοστών ανέρχεται στο 1%.

Πέρα όμως από τα χερσαία ποσοστά υπάρχουν και οι συζητήσεις για την ακτογραμμή, μια παράμετρος που κάθε άλλο παρά αμελητέα είναι σε ένα νησί το οποίο περιτριγυρίζεται από 3 ηπείρους. Τα ποσοστά της ακτογραμμής είναι εξίσου σημαντικά με το έδαφος ωστόσο τα ποσοστά αυτά δεν είναι ανάλογα. Με βάση τον χάρτη τον οποίο κατάθεσε η ΕΚ πλευρά το 2017 το ποσοστό εδάφους υπό TK διοίκηση ήταν 28.2% αλλά το ποσοστό ακτογραμμής υπό TK διοίκηση ήταν 49.6%. Ένα τεράστιο ποσοστό σε σχέση με το αντίστοιχο έδαφος. Η ακτογραμμή δεν λαμβάνεται υπόψιν άμεσα στο σύστημα μας αλλά έμμεσα με το ότι οι πράκτορες μας μπορούν να δεχτούν λιγότερο ποσοστό γης το οποίο να έχει περισσότερο ποσοστό ακτογραμμής.

Βέβαια, η TK πλευρά έχει ήδη στην κατοχή της το 36.2%, ένα σημαντικό ερώτημα που προκύπτει, αφού δεν έχουμε αναλύσει όλες τις πτυχές του κυπριακού, είναι το γιατί να συμβιβαστεί με χαμηλότερο από αυτό ποσοστό. Υπάρχουν διάφοροι λόγοι και εδώ θα αναφέρουμε τους 2 ίσως πιο σημαντικούς. Καταρχάς, τα εδάφη τα οποία έχει στην κατοχή της η TK πλευρά, δεν αναγνωρίζονται από κανένα διεθνή οργανισμό ή κράτος εκτός από την Τουρκία, με ό,τι αυτό συνεπάγεται. Επίσης, η TK πλευρά διεκδικεί ποσοστά διακυβέρνησης του νησιού τα οποία τώρα δεν κατέχει. Συνεπώς, γίνεται εύκολα αντιληπτό πως πρέπει να υπάρξει κάποιος συμβιβασμός στο εδαφικό έτσι ώστε να αποκτήσει περισσότερα σε άλλους τομείς.

Τέλος, λόγω του ότι η συζήτηση του 2017 είναι η πιο πρόσφατη συζήτηση με αριθμούς, και από τις 2 πλευρές, που έχει γίνει για την εδαφική πτυχή του Κυπριακού προβλήματος, τα δεδομένα τα οποία θα χρησιμοποιήσουμε σε αυτή την εργασία ως επιθυμητά από την κάθε πλευρά θα αφορούν τα όσα συζητήθηκαν στην Ελβετία. Αυτά τα ποσοστά (ποσοστό ικανοποίησης για τον κάθε πράκτορα) μπορούν να αλλάξουν από το parameters.txt file.

4.3 Αντιστοίχιση του Κυπριακού προβλήματος με το σύστημα

Αρχικά, στην εργασία αυτή θα ασχοληθούμε εξ ολοκλήρου με το εδαφικό και θα θεωρήσουμε πως έχουν γίνει κάποιες παραχωρήσεις από την ΕΚ πλευρά προς την ΤΚ για την διακυβέρνηση του νησιού. Έτσι, στο σύστημά μας, οι 2 πλευρές αρχίζουν από το 0 όσο αφορά το εδαφικό και η ΤΚ δεν αρνείται να δώσει κάποιο ποσοστό αφού θεωρητικά έχει πάρει ποσοστά διακυβέρνησης.

Εμβαθύνοντας περαιτέρω στο σύστημα, το διαθέσιμο ποσό το οποίο θα έχουν οι 2 πράκτορες στον πρώτο γύρο των διαπραγματεύσεων θα είναι όσο και το ποσοστό που είναι υπό κατοχή σήμερα, δηλαδή 36.2%. Όπως εξηγήσαμε και πιο πριν, για ευκολία ορίσαμε μια μονάδα εδάφους στο σύστημα να αντιστοιχεί με 0.1% εδάφους στην πραγματικότητα. Συνεπώς, το αρχικό Pot θα ισούται με 362 μονάδες.

Mapping	
Initial Pot	362

Πίνακας 4.3.1

Πίνακας αρχικού ποσού

Ο πίνακας παρουσιάζει το αρχικό ποσό για το οποίο διαπραγματεύονται οι 2 πράκτορες.

Συνεχίζοντας, έχουμε τα επιθυμητά ποσοστά 8% και 29.2% από την ΕΚ και ΤΚ πλευρά αντίστοιχα. Με αυτά τα ποσοστά δεν μπορεί να υπάρξει συμβιβασμός. Ορίσαμε το επιθυμητό ποσοστό της ΕΚ πλευράς στο 6.8% και το επιθυμητό ποσοστό της ΤΚ πλευράς στο 27.8%. Αυτή η επιλογή έγινε με το κριτήριο ότι υπάρχουν περιοχές οι οποίες είναι και για τις 2 πλευρές ζωτικής σημασίας (όπως για παράδειγμα η Μόρφου). Οι δύο πλευρές θα μπορούσαν να αποδεχθούν μικρότερο ποσοστό με σκοπό να διασφαλίσουν καίρια εδάφη. Επίσης, ένας άλλος λόγος για τον οποίο οι δύο πλευρές θα δέχονταν μικρότερο ποσοστό εδάφους είναι το ποσοστό της ακτογραμμής, διότι παρόλο που μπορεί να μειωθεί το ποσοστό του εδάφους, αυτό δεν είναι αντίστοιχο με το ποσοστό της ακτογραμμής, η οποία είναι εξίσου σημαντική. Έτσι, όσο αφορά το σύστημά μας, έχουμε 66 μονάδες επιθυμητό αποτέλεσμα από την ΕΚ πλευρά και 278 μονάδες επιθυμητό αποτέλεσμα από την ΤΚ πλευρά.

Mapping	
Initial Pot	362
EK satisfied	66
TK satisfied	278

Πίνακας 4.3.2

Πίνακας αρχικού ποσού και μονάδων ικανοποίησης

Ο πίνακας αυτός παρουσιάζει τις μονάδες εδάφους τις οποίες επιδιώκουν οι δύο πράκτορες να καταλήξουν στα χέρια τους έτσι ώστε να είναι ικανοποιημένοι. Ο ΕΚ πράκτορας επιθυμεί 66 από τις 362 συνολικές μονάδες για να είναι ικανοποιημένος και ο ΤΚ πράκτορας επιδιώκει 278 από τις 362 μονάδες για να είναι ικανοποιημένος.

Ο αρχικός πίνακας καταστάσεων για τους δύο πράκτορες είναι ο ίδιος, λόγω του ότι θεωρούμε πως η ΤΚ πλευρά έχει ήδη συμφωνήσει και πάρει ποσοστά διακυβέρνησης, και παρουσιάζεται πιο κάτω στον πίνακα 4.3.3.

State	Action/Decision		Value function
Available Area/Pot	Offer		
362	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0
361	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0
360	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0
359	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0
.	.	.	.

.	.	.	.
2	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0
1	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0
0	>= satisfied	Accept	0
		Refuse	0
	< satisfied	Accept	0
		Refuse	0

Πίνακας 4.3.3

Πίνακας Καταστάσεων και ενεργειών

Στον πίνακα αυτό απεικονίζονται οι καταστάσεις που μπορεί να βρεθεί ένας παίκτης, από την 362^η κατάσταση μέχρι και την κατάσταση 0. Σε κάθε κατάσταση βλέπουμε τις επιλογές ενεργειών που έχει ο πράκτορας και την τιμή της συνάρτησης αξίας.

Όπως δείχνει και ο πιο πάνω πίνακας, όλες οι συναρτήσεις αξίας των 2 πρακτόρων αρχικά είναι 0 και ο πράκτορας έχει όλες τις πιθανές ενέργειες κάθε κατάστασης στην διάθεση του.

Στο σύστημά μας με το τέλος κάθε γύρου μπορούμε να έχουμε 4 καταστάσεις. Οι καταστάσεις αυτές είναι οι εξής:

1. Συμβιβασμός,
2. Μη συμβιβασμός
3. Νίκη EK
4. Νίκη TK.

Για να επιτύχουμε συμβιβασμό θα πρέπει κάποιος πράκτορας να αποδεχθεί την προσφορά του άλλου πράκτορα, η οποία προσφορά θα αφήνει τον κάθε ένα με ποσοστό μεγαλύτερο ή ίσο με το ποσοστό ικανοποίησης του πράκτορα (πίνακας 4.3.2). Μη

συμβιβασμός μπορεί να υπάρξει στην περίπτωση που είτε ο ένας πράκτορας θα αποδεχθεί προσφορά η οποία αφήνει και τους 2 πράκτορες με ποσοστό μικρότερο από αυτό που τους ικανοποιεί είτε στην περίπτωση όπου οι πράκτορες δεν αποδέχονται καμία προσφορά και έτσι το Pot εξαντλείται και δεν μπορεί να υπάρξει πλέον συμφωνία. Νίκη για την μία πλευρά υπάρχει όταν ο πράκτορας αποδεχθεί μια προσφορά η οποία είναι μικρότερη από το ποσοστό που τον ικανοποιεί αλλά το ποσοστό που μένει για τον αντίπαλο είναι ικανοποιητικό για αυτόν και αντίστροφα. Όλα αυτά παρουσιάζονται συνοπτικά στον πίνακα **4.3.4**

Ενδεχόμενο	Ποσό ΕΚ (σε μονάδες συστήματος)	Ποσό ΤΚ (σε μονάδες συστήματος)
Συμβιβασμός	>=66	>=278
Μη-Συμβιβασμός	<66	<278
Νίκη ΕΚ	>=66	<278
Νίκη ΤΚ	<66	>=278

Πίνακας 4.3.4

Πίνακας Αποτελεσμάτων

Ο πίνακας αυτός μας δείχνει το αποτέλεσμα μιας επανάληψής ανάλογα με τις μονάδες εδάφους που έχει στην κατοχή του ο κάθε πράκτορας στο τέλος της επανάληψης. Όταν και οι δύο πράκτορες έχουν στην κατοχή τους ποσό μεγαλύτερο ή ίσο με αυτό που τους ικανοποιεί, δηλαδή 66 για τον ΕΚ και 278 για τον ΤΚ, τότε το αποτέλεσμα είναι συμβιβασμός. Όταν κανένας από τους δύο πράκτορες δεν πήρε ποσό τουλάχιστον ίσο με αυτό που τον ικανοποιεί έχουμε μη-συμβιβασμό. Τέλος, όταν ο ένας πράκτορας πάρει μονάδες που να είναι ίσες ή περισσότερες με αυτές που τον ικανοποιούν αλλά ο άλλος πράκτορας πάρει λιγότερες από τις μονάδες που τον ικανοποιούν τότε έχουμε μονόπλευρη νίκη, είτε για τον ΕΚ είτε για τον ΤΚ.

4.4 Αντιστοίχιση Εξωτερικών Παραγόντων με το σύστημα

Οι εξωτερικοί παράγοντες αποτελούν τους έμμεσους παίκτες του συστήματος. Δεν υλοποιήθηκαν ως παίκτες αλλά, όπως εξηγήθηκε πιο πάνω, με την ενεργοποίησή τους έχουν την δυνατότητα να ενισχύσουν ή να τιμωρήσουν τους δύο πράκτορες ανάλογα με τη «συμπεριφορά» τους. Η υλοποίηση των Εξωτερικών Παραγόντων έγινε ως μια κλάση η οποία με είσοδο την κατάσταση του συστήματος, την προσφορά του ενός πράκτορα

και την απάντηση του άλλου πράκτορα υπολογίζει και επιστρέφει το ολικό ενισχυτικό σήμα, και από τους 5 εξωτερικούς παράγοντες. Περισσότερες λεπτομέρειες θα αναφερθούν στο υποκεφάλαιο 5.4.

Αρχικά, οι πρώτοι 2 εξωτερικοί παράγοντες, αυτοί που έχουν μονόπλευρη θέση, είναι η Ελλάδα και η Τουρκία. Οι δύο αυτές χώρες στο σύστημα μας δίνουν ενισχυτικό σήμα μόνο στον πράκτορα από τον οποίο έχουν συμφέρον, τον αμείβουν αν η απόφαση είναι υπέρ τους ή τον τιμωρούν αν είναι κατά τους. Στους πίνακες 4.4.1 και 4.4.2 φαίνονται τα ενισχυτικά σήματα που προσφέρει Ελλάδα και Τουρκία αντίστοιχα. Στις παρενθέσεις φαίνεται η ποσότητα του Ενισχυτικού σήματος που δίνεται στην κάθε περίπτωση. Η ποσότητα που δίνεται ως επιβράβευση είναι μικρότερη από αυτή που δίνεται ως τιμωρία διότι η κάθε χώρα μπορεί να τιμωρήσει κατά πολύ περισσότερο τον πράκτορα από τον οποίο έχει συμφέρον παρά να τον επιβραβεύσει.

Ενισχυτικό σήμα Ελλάδας

Πράκτορας που αποφασίζει	Προσφερόμενες μονάδες για πράκτορα που αποφασίζει			
	>=satisfied		<satisfied	
	Αποδοχή προσφοράς	Απόρριψη προσφοράς	Αποδοχή προσφοράς	Απόρριψη προσφοράς
EK	Reward (1)	Penalty (-5)	Penalty (-5)	Reward (1)
TK	N/A	N/A	N/A	N/A

Πίνακας 4.4.1

Πίνακας Ενισχυτικού σήματος Ελλάδας

Στον πίνακα αυτό βλέπουμε πότε η Ελλάδα προσφέρει επιβράβευση στον EK πράκτορα και πότε τιμωρία. Όταν αποδέχεται προσφορά που τον ικανοποιεί ή όταν απορρίπτει προσφορά που δεν τον ικανοποιεί τότε δίνεται επιβράβευση. Στις δύο περιπτώσεις κατά τις οποίες απορρίπτει προσφορά η οποία θα τον άφηνε ικανοποιημένο ή αποδέχεται προσφορά που δεν τον αφήνει ικανοποιημένο δίνεται τιμωρία. Το ποσό της τιμωρίας είναι μεγαλύτερο από αυτό της επιβράβευσης επειδή η Ελλάδα μπορεί να τιμωρήσει αρκετά τον EK πράκτορα αλλά όχι τόσο να τον επιβραβεύσει.

Ενισχυτικό σήμα Τουρκίας

Πράκτορας που αποφασίζει		Προσφερόμενες μονάδες για πράκτορα που αποφασίζει		
	>=satisfied		<satisfied	
	Αποδοχή προσφοράς	Απόρριψη προσφοράς	Αποδοχή προσφοράς	Απόρριψη προσφοράς
EK	N/A	N/A	N/A	N/A
TK	Reward (1)	Penalty (-5)	Penalty (-5)	Reward (1)

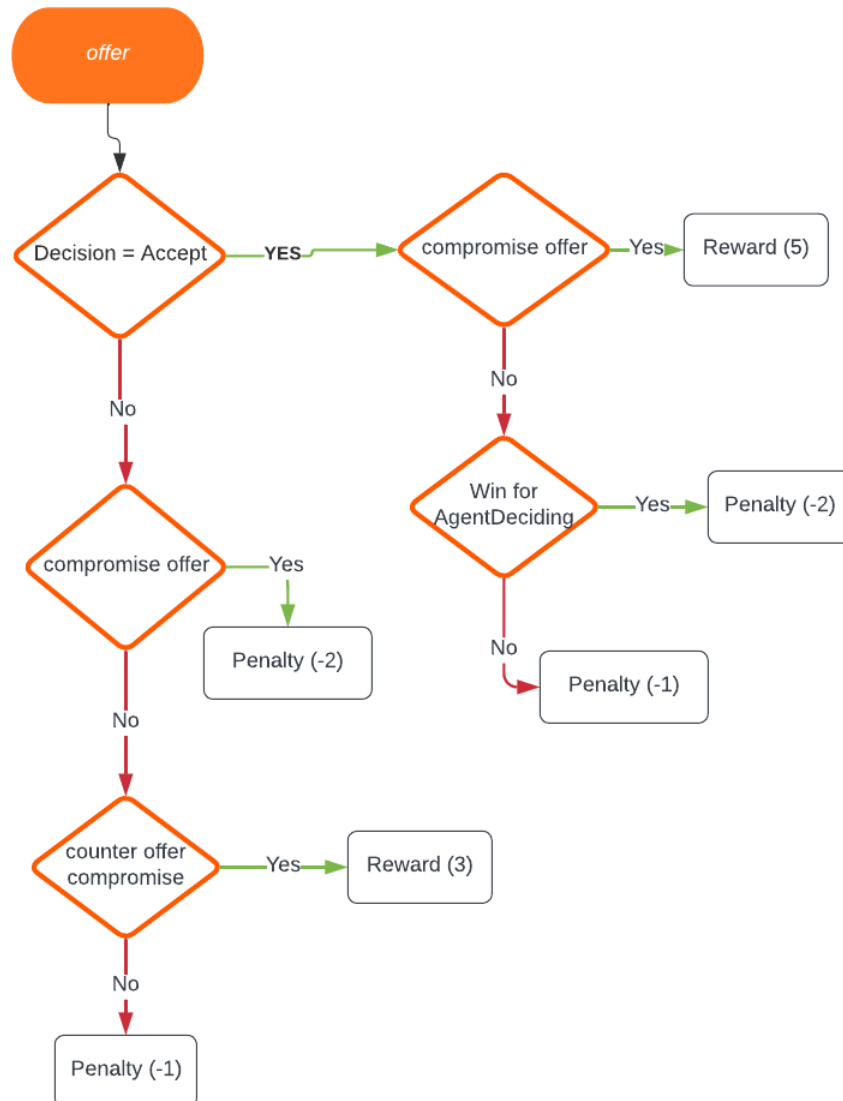
Πίνακας 4.4.2

Πίνακας Ενισχυτικού σήματος Τουρκίας

Στον πίνακα αυτό βλέπουμε πότε η Τουρκία προσφέρει επιβράβευση στον TK πράκτορα και πότε τιμωρία. Όταν αποδέχεται προσφορά που τον ικανοποιεί ή όταν απορρίπτει προσφορά που δεν τον ικανοποιεί τότε δίνεται επιβράβευση. Στις δύο περιπτώσεις κατά τις οποίες απορρίπτει προσφορά η οποία θα τον αφήνε ικανοποιημένο ή αποδέχεται προσφορά που δεν τον αφήνει ικανοποιημένο δίνεται τιμωρία. Το ποσό της τιμωρίας είναι μεγαλύτερο από αυτό της επιβράβευσης επειδή η Τουρκία μπορεί να τιμωρήσει αρκετά τον TK πράκτορα αλλά όχι τόσο να τον επιβραβεύσει.

Έπειτα, τα Ηνωμένα Έθνη, με βάση τις αρχές και τις θέσεις του οργανισμού στις διάφορες συνομιλίες που έγιναν, επιδιώκουν τον συμβιβασμό και την ειρήνη μεταξύ των δύο πλευρών. Χρησιμοποιώντας αυτές τις θέσεις ορίσαμε το ενισχυτικό σήμα του οργανισμού αυτού να επιβραβεύει τους παίκτες όταν αυτοί φτάσουν σε συμβιβασμό, να τιμωρεί τους παίκτες όταν αποδεχτούν πρόταση η οποία συμφέρει μόνο τους ίδιους και να τιμωρεί τους παίκτες όταν απορρίψουν προσφορά συμβιβασμού ή όταν απορρίψουν προσφορά και κάνουν αντιπροσφορά η οποία να μην οδηγεί σε συμβιβασμό. Στην περίπτωση όμως που ο πράκτορας απορρίψει πρόταση η οποία δεν οδηγούσε σε συμβιβασμό αλλά κάνει μια αντιπρόταση που οδηγεί σε συμβιβασμό ο πράκτορας αυτός επιβραβεύεται.

UN Reinforcement Signal



Διάγραμμα 4.4.3

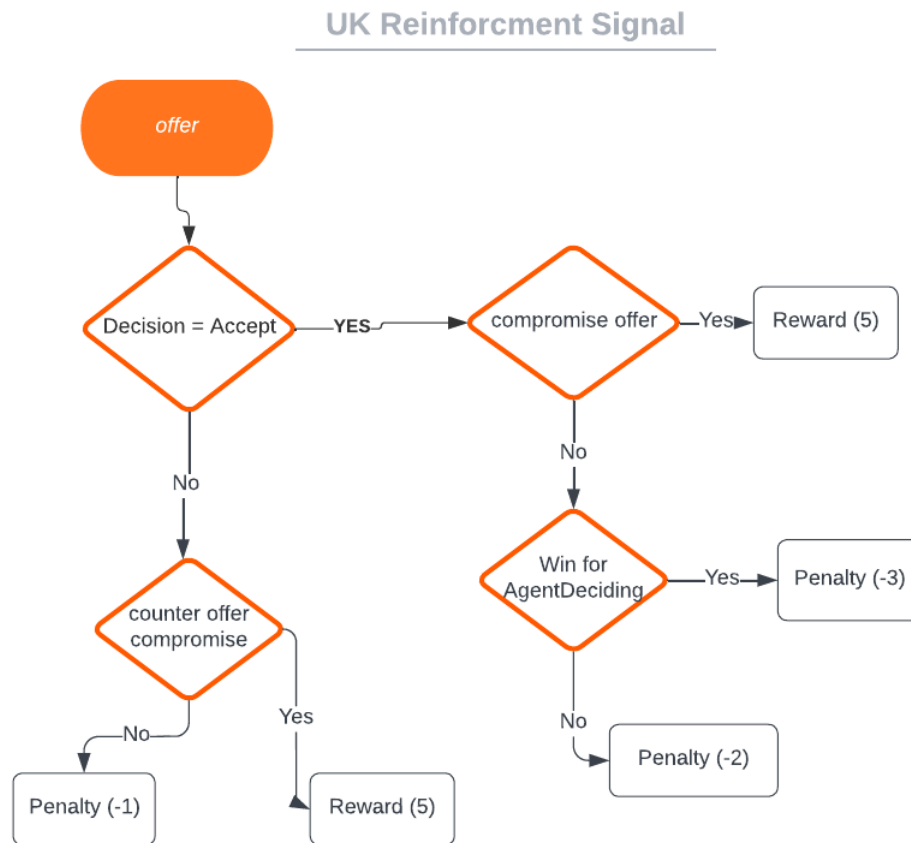
Πίνακας Ενισχυτικού σήματος Οργανισμού Ηνωμένων Εθνών

Στον πίνακα αυτό βλέπουμε πότε ο ΟΗΕ επιβραβεύει ή τιμωρεί τους πράκτορες μας ανάλογα με την απόφασή τους. Όταν ο πράκτορας αποδεχθεί μια προσφορά η οποία επιφέρει και στους δύο πράκτορες ικανοποίηση, άρα οδηγεί σε συμβιβασμό, ο πράκτορας επιβραβεύεται με ένα θετικό σήμα 5. Αν ο πράκτορας αποδέχτηκε προσφορά η οποία επέφερε μονόπλευρη νίκη τότε τιμωρείται με -2. Στην περίπτωση όπου ο πράκτορας αρνήθηκε προσφορά η οποία ήταν προσφορά συμβιβασμού τιμωρείται με -2. Αν η προσφορά αυτή δεν ήταν συμβιβασμού και ο πράκτορας έκανε αντιπροσφορά συμβιβασμού επιβραβεύεται με θετικό ενισχυτικό σήμα 3. Αν τώρα έκανε αντιπροσφορά που πάλι δεν οδηγεί σε συμβιβασμό τιμωρείται με -1.

Το Διάγραμμα Ροής 4.4.3 παρουσιάζει όλες τις πιθανές περιπτώσεις για τον πράκτορα που αποφασίζει και το Ενισχυτικό σήμα που παράγει ο εξωτερικός παράγοντας Ο.Η.Ε., οι τιμές φαίνονται στις παρενθέσεις. Ο πράκτορας τιμωρείται στις περιπτώσεις όπου αρνείται πρόταση συμβιβασμού με -2, ή αν αρνήθηκε πρόταση η οποία δεν οδηγούσε σε συμβιβασμό και έκανε αντιπρόταση που πάλι δεν οδηγεί σε συμβιβασμό (-1). Αν αρνήθηκε πρόταση που δεν οδηγούσε σε συμβιβασμό αλλά έκανε πρόταση συμβιβασμού τότε επιβραβεύεται με ενισχυτικό σήμα = 3. Στην περίπτωση όπου ο πράκτορας αποδέχτηκε προσφορά συμβιβασμού επιβραβεύεται με τιμή 5. Αν όμως η αποδοχή της προσφοράς τον οδήγησε σε μονόπλευρη νίκη για αυτόν τιμωρείται με -2, αν οδήγησε σε μονόπλευρη νίκη για τον αντίπαλο ή μη συμβιβασμό τιμωρείται με -1. Όλα αυτά βασίζονται στην σχέση μεταξύ τους και τις επιδιώξεις και δυνατότητες του ΟΗΕ. Δηλαδή έχει περισσότερες δυνατότητες να επιβραβεύσει τους δύο πράκτορες παρά να τους τιμωρήσει.

Στην συνέχεια, έχουμε το ΗΒ το οποίο την παρούσα στιγμή (2022) διαθέτει 2 στρατιωτικές βάσεις στο νησί μια στη χερσόνησο Ακρωτήρι και μια στην Δεκέλεια, η οποίες καλύπτουν περίπου το 3% της συνολικής γης της Κύπρου. Το ΗΒ όμως δεν έχει υπό την κυριαρχία του καθόλου ΑΟΖ, γεγονός για το οποίο υπήρξαν μερικές συζητήσεις κυρίως στο σχέδιο Ανάν 5. Το ΗΒ ως εγγυήτρια δύναμη και μέσα από τις συζητήσεις φαίνεται να επιθυμεί μια λύση συμβιβασμού και αυτό. Το βέβαιο είναι πως δεν θέλει επιπλέον διαμάχες μεταξύ των 2 πλευρών, αφού δεν θέλει να υπάρξει οποιαδήποτε πιθανότητα να διαφοροποιηθεί κάτι όσο αφορά την κυριαρχία του στις βάσεις που διατηρεί. Είναι εμφανές πως με τις συζητήσεις για το Κυπριακό και για το 36.2% του εδάφους του νησιού τα σχετικά με αυτό μικρά κομμάτια γης που υπάγονται κάτω από το ΗΒ έρχονται σε δεύτερη μοίρα και δεν υπάρχει άμεσος λόγος για συζητήσεις. Επιπρόσθετα, το ΗΒ προσφέρθηκε (στο Ανάν 5) να παραχωρήσει, σε περίπτωση λύσης, ένα αρκετά σημαντικό ποσοστό από τις στρατιωτικές βάσεις που διατηρεί στο νησί, στην ενωμένη Κυπριακή Δημοκρατία. Η προθυμία αυτή φαίνεται να ισχύει μέχρι και σήμερα χωρίς μάλιστα τις απαιτήσεις για κομμάτι της ΑΟΖ. Αυτό αποτελεί πραγματική επιβράβευση προς τους 2 πράκτορες σε περίπτωση συμβιβασμού. Συνεπώς, σε γενικές γραμμές επιθυμεί μια λύση συμβιβασμού χωρίς να έχει ιδιαίτερο συμφέρον από αυτήν

γι' αυτό και δεν την επιδιώκει ιδιαίτερα. Βασικότερη επιθυμία του είναι να μην υπάρξει μια μονόπλευρη λύση και να εμφανιστούν μεγαλύτερες εντάσεις.



Διάγραμμα 4.4.4

Πίνακας Ενισχυτικού σήματος Ηνωμένου Βασιλείου

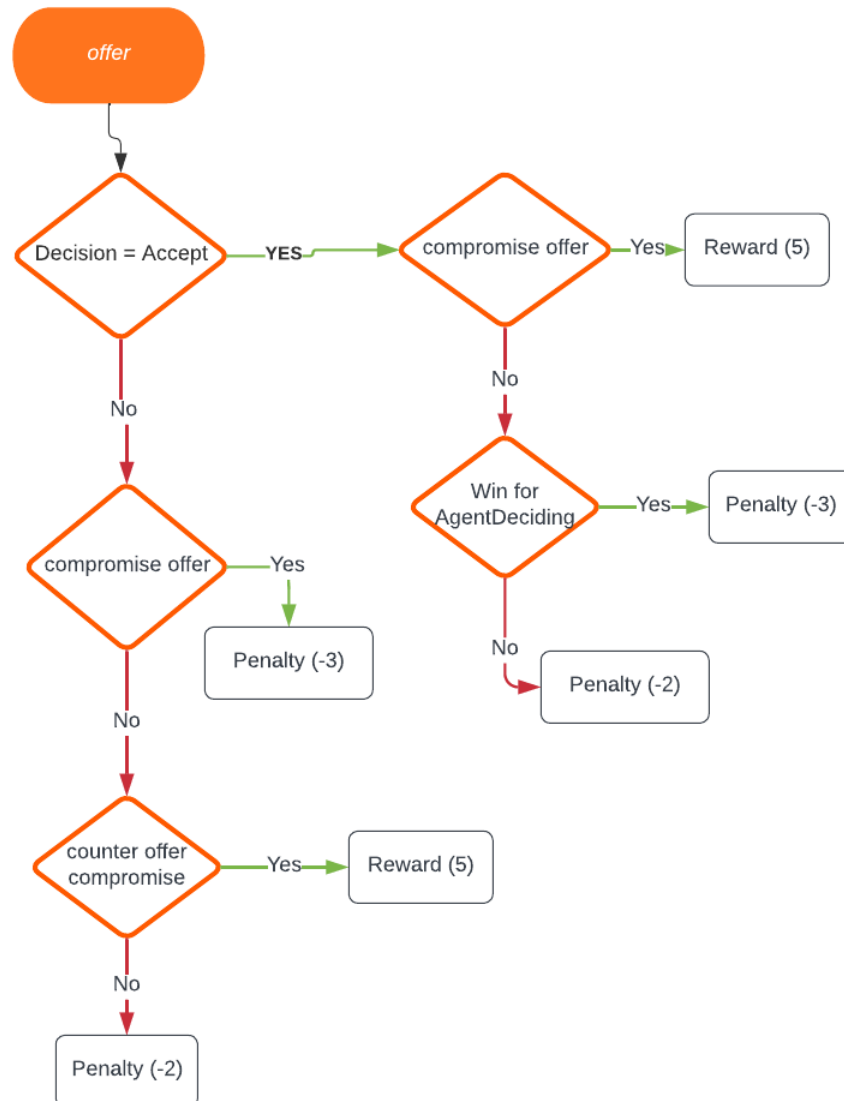
Στον πίνακα αυτό βλέπουμε πότε το HB επιβραβεύει ή τιμωρεί τους πράκτορες μας ανάλογα με την απόφασή τους. Όταν ο πράκτορας αποδεχθεί μια προσφορά η οποία επιφέρει και στους δύο πράκτορες ικανοποίηση, άρα οδηγεί σε συμβιβασμό, ο πράκτορας επιβραβεύεται με ένα θετικό σήμα 5. Αν ο πράκτορας αποδέχτηκε προσφορά η οποία επέφερε μονόπλευρη νίκη τότε τιμωρείται με -3. Στην περίπτωση όπου ο πράκτορας αρνήθηκε προσφορά αν η αντιπροσφορά που έκανε είναι προσφορά συμβιβασμού επιβραβεύεται με θετικό ενισχυτικό σήμα 5. Αν τώρα έκανε αντιπροσφορά που πάλι δεν οδηγεί σε συμβιβασμό τιμωρείται με -1.

Το Διάγραμμα Ροής 4.4.4 παρουσιάζει όλες τις πιθανές περιπτώσεις για τον πράκτορα που αποφασίζει και το ενισχυτικό σήμα που παράγει ο εξωτερικός παράγοντας: HB. Βλέπουμε πως εάν ο πράκτορας αποδεχθεί μια προσφορά η οποία οδήγησε σε

συμβιβασμό θα επιβραβευθεί με 5, αν όμως αποδεχθεί μια προσφορά η οποία οδηγεί σε νίκη για τον ένα πράκτορα ή και μη-συμβιβασμό θα τιμωρηθεί με -3 και -2 αντίστοιχα. Αν τώρα ο πράκτορας αρνηθεί την προσφορά και κάνει αντιπροσφορά για συμβιβασμό θα επιβραβευθεί με 5, αν όχι θα τιμωρηθεί με -1. Οι επιβραβεύσεις του HB έχουν μεγαλύτερες τιμές λόγω του ότι υπάρχουν συγκεκριμένη ανταμοιβή και στους 2 παίκτες σε περίπτωση που υπάρξει λύση συμβιβασμού. Το HB στο σχέδιο Ανάν 5 εξέφρασε την προθυμία του να προσφέρει ένα ποσοστό από τις στρατιωτικές βάσεις που κατέχει στην Ενωμένη Κυπριακή Δημοκρατία.

Τέλος, οι ΗΠΑ έχουν σημαντικό ρόλο να διαδραματίσουν στο πρόβλημα αφού μπορούν να επιφέρουν κυρώσεις τόσο στους 2 κύριους πράκτορες μας, ΤΚ και ΕΚ, αλλά κυρίως στις 2 εξωτερικές πηγές πίσω από αυτούς, Τουρκία και Ελλάδα αντίστοιχα. Τα συμφέροντα των ΗΠΑ φαίνεται να κλείνουν προς τον συμβιβασμό κυρίως διότι με την ειρήνη στο νησί, οι ΗΠΑ μπορούν να έχουν πρόσβαση σε όλο το νησί και να έχουν τη δυνατότητα να χρησιμοποιήσουν την τόσο ευνοϊκή θέση του νησιού για να ελέγξουν τη γύρω περιοχή. Οι ΗΠΑ φαίνεται να στηρίζουν και να συνεργάζονται με το νόμιμο και αναγνωρισμένο καθεστώς διακυβέρνησης της Κύπρου το οποίο αυτή τη στιγμή απαρτίζεται μόνο από ΕΚ. Για τις ΗΠΑ θα ήταν πιο συμφέρον να συμμετέχει και η Τουρκία στο καθεστώς, ως μέλος του ΝΑΤΟ γι' αυτό και τείνουν προς τον συμβιβασμό. Βέβαια, το τελευταίο διάστημα οι σχέσεις μεταξύ ΗΠΑ και Τουρκίας φαίνεται να κλονίζονται. Λαμβάνοντας και αυτό υπόψιν όταν ο ΤΚ πράκτορας παίρνει αποφάσεις που δεν βοηθούν να φτάσουν σε συμβιβασμό τιμωρείται περισσότερο.

USA Reinforcment Signal



Διάγραμμα 4.4.5

Πίνακας Ενισχυτικού σήματος Ηνωμένων Πολιτειών Αμερικής

Στον πίνακα αυτό βλέπουμε πότε οι ΗΠΑ επιβραβεύουν ή τιμωρούν τους πράκτορες μας ανάλογα με την απόφασή τους. Όταν ο πράκτορας αποδεχθεί μια προσφορά η οποία επιφέρει και στους δύο πράκτορες ικανοποίηση, άρα οδηγεί σε συμβιβασμό, ο πράκτορας επιβραβεύεται με ένα θετικό σήμα 5. Αν ο πράκτορας αποδέχτηκε προσφορά η οποία επέφερε μονόπλευρη νίκη τότε τιμωρείται με -3. Στην περίπτωση όπου ο πράκτορας αρνήθηκε προσφορά η οποία ήταν προσφορά συμβιβασμού τιμωρείται με -3. Αν η προσφορά αυτή δεν ήταν συμβιβασμού και ο πράκτορας έκανε αντιπροσφορά συμβιβασμού επιβραβεύεται με θετικό ενισχυτικό σήμα 5. Αν τώρα έκανε αντιπροσφορά που πάλι δεν οδηγεί σε συμβιβασμό τιμωρείται με -2.

Το Διάγραμμα Ροής 4.4.5 παρουσιάζει όλες τις πιθανές περιπτώσεις για τον πράκτορα που αποφασίζει και το ενισχυτικό σήμα που παράγει ο εξωτερικός παράγοντα ΗΠΑ. Ο πράκτορας τιμωρείται στις περιπτώσεις όπου αρνείται πρόταση συμβιβασμού με -3, ή αν αρνήθηκε πρόταση η οποία δεν οδηγούσε σε συμβιβασμό και έκανε αντιπρόταση που πάλι δεν οδηγεί σε συμβιβασμό (-2). Αν αρνήθηκε πρόταση που δεν οδηγούσε σε συμβιβασμό αλλά έκανε πρόταση συμβιβασμού τότε επιβραβεύεται με ενισχυτικό σήμα ίσο με 5. Στην περίπτωση όπου ο πράκτορας αποδέχτηκε προσφορά συμβιβασμού επιβραβεύεται με τιμή 5. Αν όμως η αποδοχή της προσφοράς τον οδήγησε σε μονόπλευρη νίκη για αυτόν τιμωρείται με -3, αν οδήγησε σε μονόπλευρη νίκη για τον αντίπαλο ή μη συμβιβασμό τιμωρείται με -2. Όλα αυτά βασίζονται στην σχέση μεταξύ τους και τις επιδιώξεις και δυνατότητες των ΗΠΑ, είναι φανερό πως οι ΗΠΑ μπορούν να επιβάλουν μεγαλύτερες κυρώσεις ή επιβραβεύσεις στους δύο πράκτορες για αυτό και η διαφορά στα σήματα με τα σήματα του ΟΗΕ.

Κεφάλαιο 5

Αποτελέσματα και Συζήτηση

5.1 Εισαγωγή	52
5.2 Προσομοίωση με την χρήση Q-Learning agent	53
5.3 Προσομοίωση με την χρήση SARSA agent	61
5.4 Προσθήκη Άδηλων Πηγών Ενίσχυσης	64
5.5 Προσθήκη Επιπρόσθετης Αμοιβής	68

5.1 Εισαγωγή

Στο κεφάλαιο αυτό θα αναλύσουμε διάφορα αποτελέσματα με διαφορετικές παραμέτρους χρησιμοποιώντας τους δύο πράκτορες μας Q-learning και SARSA στην προσομοίωση μας. Οι δύο αυτοί πράκτορες χρησιμοποιούν την e-greedy τεχνική και κρατάμε σταθερό το ϵ . Οι συνδυασμοί των παραμέτρων είναι πάρα πολλοί οπότεν δοκιμάσαμε διάφορους και θα παρουσιάσουμε αυτούς που έχουν κάτι διαφορετικό να μας προσφέρουν.

Μετά την ανάλυση μερικών αποτελεσμάτων χωρίς την χρήση των ΑΠΕ θα συγκρίνουμε αποτελέσματα χωρίς αλλά και με την χρήση των ΑΠΕ. Μετά από διάφορες δοκιμές παρατηρήσαμε πως μετά τις 500 εποχές τα αποτελέσματα σταθεροποιούνται οπότεν δεν χρειάζεται να εκτελούμε περισσότερες και έτσι στα αποτελέσματα μας χρησιμοποιούμε 500 εποχές. Για κάθε επιλογή παραμέτρων τρέξαμε το πρόγραμμα για 100 φορές και χρησιμοποιούμε τον μέσο όρο στα αποτελέσματα έτσι ώστε να έχουμε πιο σωστά αποτελέσματα.

Επιπρόσθετα, θα πρέπει να τονίσουμε πως λόγω του ότι ο ΕΚ πράκτορας έχει την δυνατότητα να κερδίσει σε πολύ μεγαλύτερο αριθμό γύρων (296 πρώτους γύρους) σε

σχέση με τον TK πράκτορα (94 πρώτους γύρους), αφού απαιτεί ένα πολύ μικρότερο ποσοστό για να είναι ικανοποιημένος. Αυτό φαίνεται στα αποτελέσματα διότι στους αρχικούς γύρους φαίνεται να ικανοποιείται κυρίως ο ΕΚ.

5.2 Προσομοίωση με την χρήση Q-Learning agent

Αρχικά, θα δούμε την επίδραση του ρυθμού μάθησης αλλά και του ρυθμού έκπτωσης στα αποτελέσματα του κάθε γύρου. Κρατάμε σταθερό τον ένα παράγοντα και μεταβάλλουμε τον άλλο. Θα παρουσιάσουμε τα πιο σημαντικά από αυτά τα αποτελέσματα.

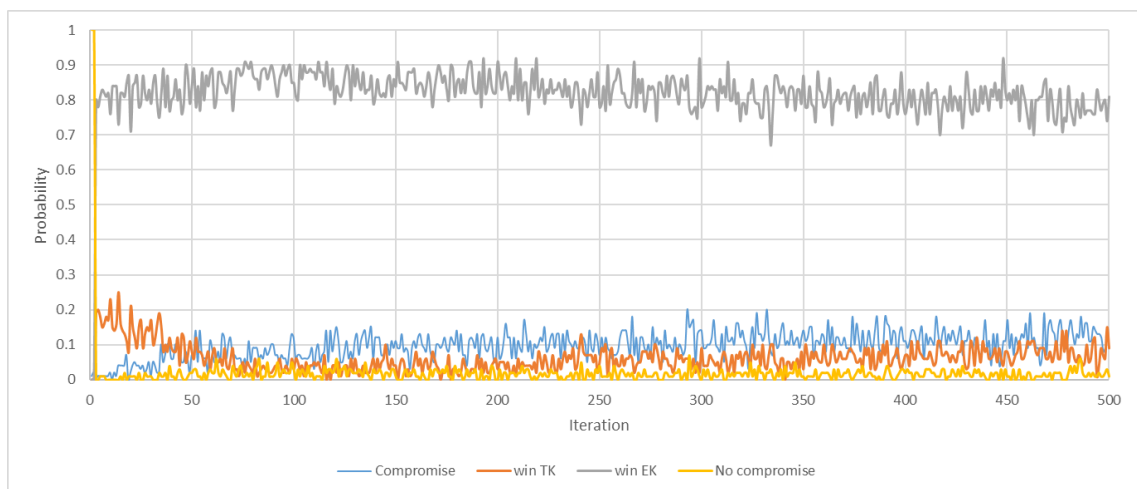
1^η Περίπτωση

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
ΕΚ	0.3	0.5
TK	0.3	0.5

Πίνακας 5.2.1

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.2.1



Σχήμα 5.2.1

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.2.1

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας δεν μαθαίνουν να συμβιβάζονται και δεν αλλάζουν ιδιαίτερα συμπεριφορά

Όπως είναι εμφανές από την πιο πάνω γραφική με τις παραμέτρους αυτές έχουμε πολύ σταθερά αποτελέσματα τα οποία μεταβάλλονται ελάχιστα με την πάροδο των επαναλήψεων. Το ποσοστό των συμβιβασμών αυξάνεται ελάχιστα και τα ποσοστά νίκης ενός πράκτορα μειώνονται ελάχιστα.

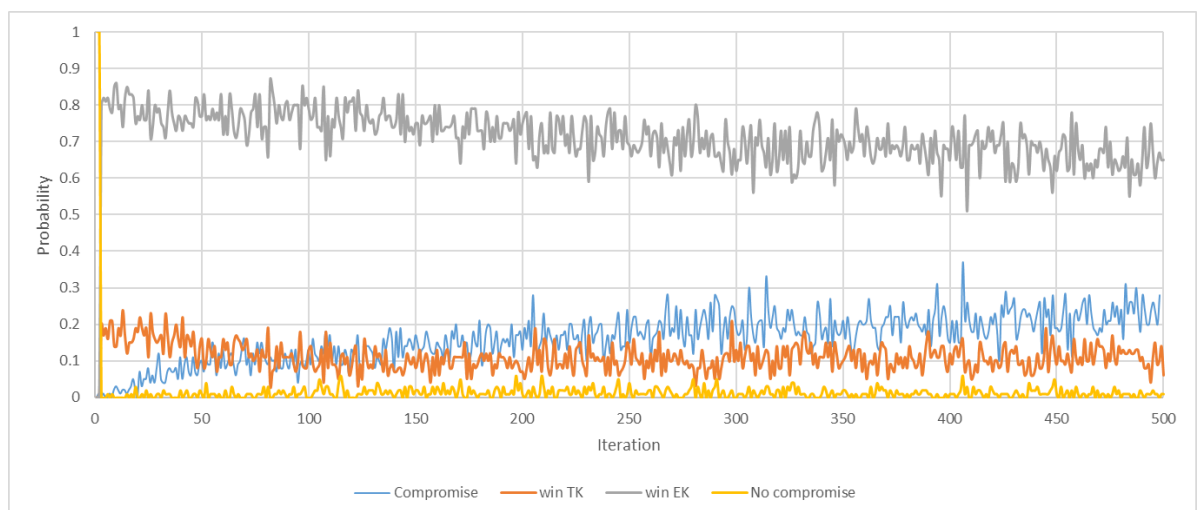
2^η Περίπτωση

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
EK	0.4	0.5
TK	0.4	0.5

Πίνακας 5.2.2

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.2.2



Σχήμα 5.2.2

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.2.2

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας μαθαίνουν ελάχιστα να συμβιβάζονται και υπάρχει μια αυξητική τάση στην πιθανότητα συμβιβασμού. Δείχνει επίσης πως οι νίκες του ΕΚ πράκτορα μειώνονται με την πάροδο των επαναλήψεων

Σε αυτό το σχήμα με τις παραμέτρους αυτές έχουμε μια μικρή σχετικά πρόοδο με την πάροδο των επαναλήψεων. Βλέπουμε τις νίκες του ΕΚ να μειώνονται και τους συμβιβασμούς να αυξάνονται και να φτάνουν ένα ποσοστό της τάξης του 30%. Φαίνεται πως με την αύξηση του ρυθμού μάθησης οι πράκτορες αρχίζουν να μαθαίνουν και έχουμε μια τάση για αύξηση του ποσοστού συμβιβασμού.

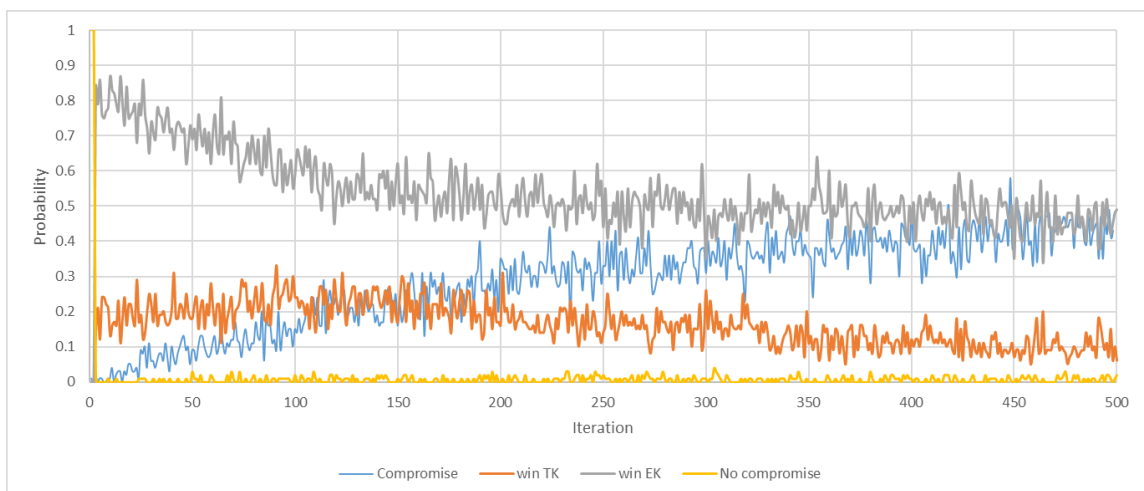
3^η Περίπτωση

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
ΕΚ	0.5	0.5
ΤΚ	0.5	0.5

Πίνακας 5.2.3

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.2.3



Σχήμα 5.2.3

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.2.3

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας εκπαιδεύονται αρκετά και μαθαίνουν να συμβιβάζονται. Το ποσοστό συμβιβασμού συγκλίνει με τις νίκες του ΕΚ πράκτορα οι οποίες και μειώνονται αρκετά από την αρχική επανάληψη.

Σε αυτό το σχήμα με τις παραμέτρους αυτές έχουμε μια αισθητή διαφορά στα αποτελέσματα. Βλέπουμε πως από τους αρχικούς πλέον γύρους οι νίκες τόσο του ΕΚ πράκτορα όσο και του ΤΚ μειώνονται και έχουμε την ανάλογη αύξηση στην πιθανότητα για συμβιβασμό που φτάνει περίπου το 45%.

Δηλαδή, με την αύξηση του ρυθμού μάθησης στο 0.5 παρατηρούμε πως το ποσοστό συμβιβασμού με το ποσοστό νικών για τον ΕΚ πράκτορα συγκλίνει με το τέλος των επαναλήψεων. Αυτό πιθανόν να οφείλεται στο ότι ο πράκτορας λαμβάνει περισσότερο υπόψιν του τις ανταμοιβές και τις τιμωρίες με αποτέλεσμα να αλλάζει αναλόγως και η «συμπεριφορά» (επιλογές) του.

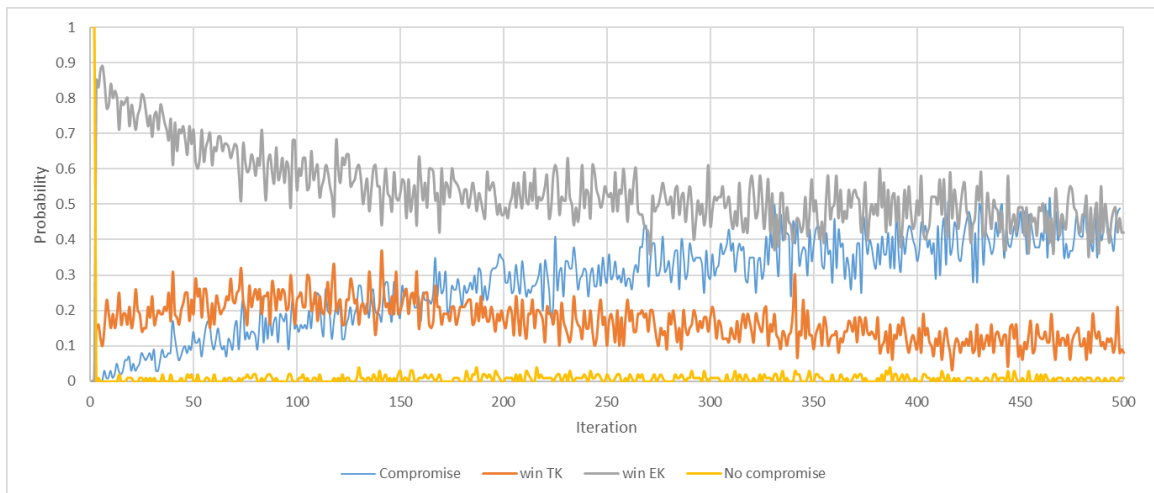
4^η Περίπτωση

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
ΕΚ	0.8	0.5
ΤΚ	0.8	0.5

Πίνακας 5.2.4

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.2.4



Σχήμα 5.2.4

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.2.4

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας εκπαιδεύονται αρκετά και μαθαίνουν να συμβιβάζονται. Το ποσοστό συμβιβασμού συγκλίνει με τις νίκες του EK πράκτορα οι οποίες και μειώνονται αρκετά από την αρχική επανάληψη.

Σε αυτή τη περίπτωση βλέπουμε πως δεν έχουμε μεγάλες διαφορές από την περίπτωση 3. Η πιθανότητα συμβιβασμού φτάνει περίπου μέχρι και το 50% μειώνοντας παράλληλα την πιθανότητα της μονόπλευρης νίκης. Ενώ στις προηγούμενες αυξήσεις του learning rate βλέπαμε το ποσοστό συμβιβασμού όλο και να αυξάνεται, μετά το 0.5 LR φαίνεται πως η αύξηση του πλέον δεν επηρεάζει το αποτέλεσμα και οι πράκτορες δεν αλλάζουν συμπεριφορά.

Έπειτα, αλλάξαμε τον παράγοντα έκπτωσης τον οποίο διατηρούμε σταθερό και μεταβάλλαμε και πάλι τον ρυθμό μάθησης.

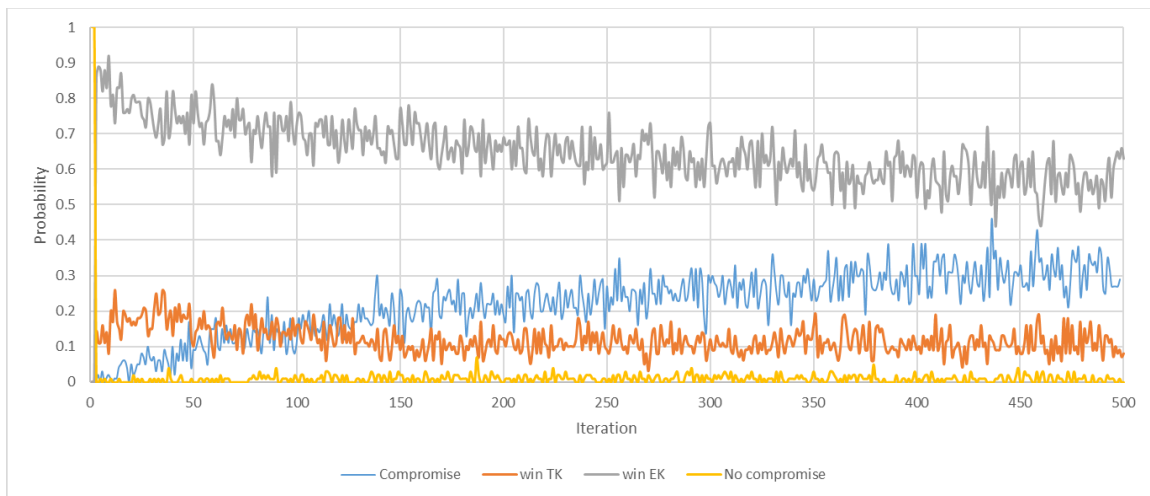
5^η Περίπτωση

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
EK	0.3	0.2
TK	0.3	0.2

Πίνακας 5.2.5

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.2.5



Σχήμα 5.2.5

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρου του πίνακα 5.2.5

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας εκπαιδεύονται αλλά όχι αρκετά, οπότεν βλέπουμε το ποσοστό συμβιβασμού να αυξάνεται χωρίς ωστόσο να φτάνει τα ψηλά ποσοστά άλλων παραμέτρων. Παράλληλα βλέπουμε τις νίκες του EK να μειώνονται αρκετά όπως και του TK μειώνονται σχεδόν στο μισό. Αυτά τα αποτελέσματα πιθανόν να οφείλονται στο χαμηλό LR των παραμέτρων αυτών αφού οι πράκτορες δεν λαμβάνουν τόσο πολύ υπόψιν τους τα ενισχυτικά σήματα και δεν είναι αρκετό για να εκπαιδευτούν επαρκώς.

Σε αυτή τη περίπτωση βλέπουμε να έχουμε πολύ καλύτερα αποτελέσματα από την περίπτωση 1 όπου και πάλι είχαμε ένα τόσο χαμηλό LR αλλά φαίνεται πως το ποσοστό των νικών του EK πράκτορα παραμένει σε υψηλά ποσοστά και δεν επιτρέπει στο ποσοστό των συμβιβασμών να ανέβει. Αυτό μας δείχνει πως όταν οι πράκτορες γίνονται πιο «μυωπικοί» δηλαδή δεν δίνουν πολλή σημασία στις μελλοντικές αξίες αλλά βλέπουν μόνο τις πιο κοντινές και παράλληλα έχουν χαμηλότερο ρυθμό μάθησης, τείνουν να μαθαίνουν καλύτερα από όταν ο ρυθμός μάθησης είναι κατά πολύ υψηλότερος από τον συντελεστή έκπτωσης.

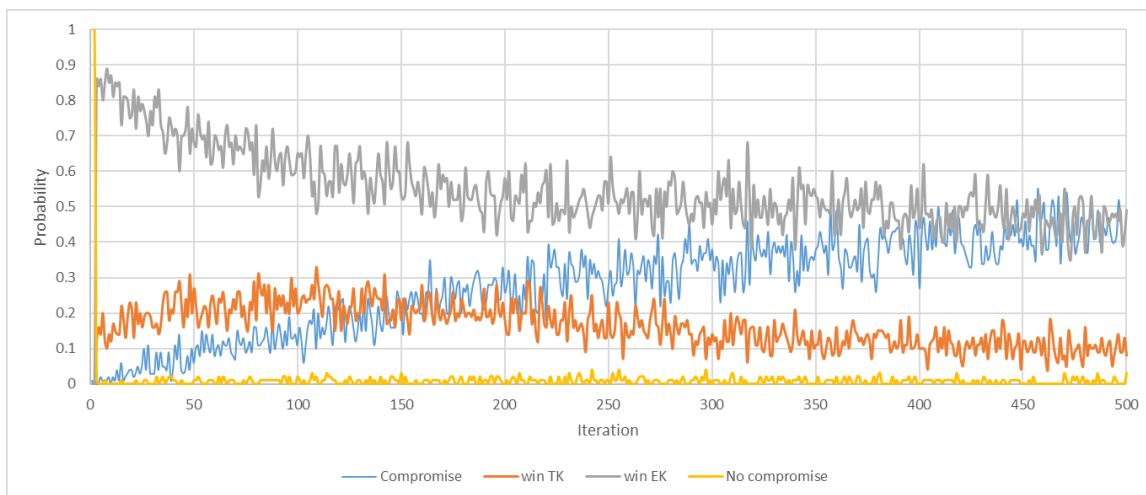
6^η Περίπτωση

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
EK	0.7	0.2
TK	0.7	0.2

Πίνακας 5.2.6

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.2.6



Σχήμα 5.2.6

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.2.6

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας εκπαιδεύονται αρκετά και μαθαίνουν να συμβιβάζονται. Το ποσοστό συμβιβασμού συγκλίνει περίπου στο 48-50% μαζί με τις νίκες του EK πράκτορα οι οποίες και μειώνονται αρκετά από την αρχική επανάληψη.

Στις υπόλοιπες δοκιμές με τον παράγοντα έκπτωσης στο 0.2 και ρυθμό μάθησης > 0.4 είχαμε παρόμοια αποτελέσματα με το σχήμα 5.2.6. Φαίνεται και πάλι να έχουμε την ίδια τάση όπως και στην 4^η περίπτωση. Ο EK πράκτορας μειώνει τις νίκες του και οι συμβιβασμοί αυξάνονται. Αυτό μας δείχνει πως σε αντίθεση με την περίπτωση 5, όταν

το learning rate είναι αρκετά υψηλό τα αποτελέσματα δεν επηρεάζονται σχεδόν καθόλου από τον συντελεστή έκπτωσης.

Μεταβάλλοντας ξανά τον παράγοντα έκπτωσης αυξάνοντας τον στο 0.7 φαίνεται πως ένας τόσο μεγάλος παράγοντας έκπτωσης δίνει ένα μεγάλο προτέρημα στον ΕΚ πράκτορα και διατηρεί ψηλά τα ποσοστά νίκης του. Στην πιο κάτω περίπτωση παρουσιάζουμε μια από αυτές τις δοκιμές αφού και οι υπόλοιπες (με διαφορετικό LR) έχουν παρόμοια αποτελέσματα.

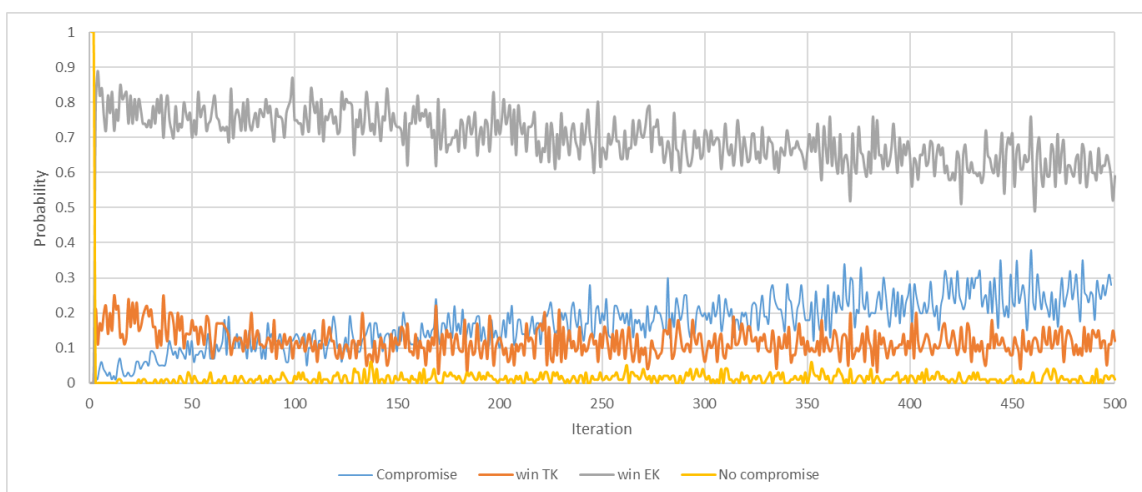
7^η Περίπτωση

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
ΕΚ	0.7	0.7
ΤΚ	0.7	0.7

Πίνακας 5.2.7

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.2.7



Σχήμα 5.2.7

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.2.7

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας δεν εκπαιδεύονται αρκετά. Το ποσοστό συμβιβασμού ανεβαίνει με την πάροδο των επαναλήψεων χωρίς ωστόσο να φτάνει σε ικανοποιητικά επίπεδα. Αυτό πιθανόν να οφείλετε στο πολύ υψηλό LR σε συνδυασμό και με τον υψηλό παράγοντα έκπτωσης ο οποίο όπως φαίνεται κάνει τους πράκτορες να βλέπουν το μελλοντικό προσωπικό τους συμφέρον.

Όπως φαίνεται στο σχήμα 5.2.7 μειώνονται ελάχιστα οι πιθανότητες μιας νίκης του ΕΚ πράκτορα και αυξάνονται λίγο το ποσοστό συμβιβασμού. Αυτό πιθανόν να οφείλεται στο ότι ο ΕΚ αφού έχει περισσότερες ευκαιρίες (περισσότερους γύρους) για να πάρει μια νίκη και αφού με ένα τόσο μεγάλο παράγοντα έκπτωσης μπορεί να «δει» πολλούς από αυτούς το εκμεταλλεύεται για να ικανοποιεί το συμφέρον του. Έτσι, βλέπουμε πως οι πράκτορες που «βλέπουν» πιο μακροπρόθεσμα, και την ίδια ώρα έχουν υψηλό ρυθμό μάθησης τείνουν να βρίσκουν τον τρόπο για να ικανοποιούν τους στόχους τους από πολύ γρήγορα τον οποίο και χρησιμοποιούν.

5.3 Προσομοίωση με την χρήση SARSA agent

Όταν χρησιμοποιήθηκε ο αλγόριθμος SARSA παρατηρήσαμε πως τα περισσότερα αποτελέσματα ήταν πολύ παρόμοια με αυτά του Q-learning, τα αποτελέσματα που είχαν αισθητή διαφορά με τον Q-learning agent είναι κυρίως αυτά με χαμηλό LR στα οποία ο SARSA πράκτορας έχει σχετικά καλύτερα αποτελέσματα όπως και σε αυτά με μεγάλο Discount Factor. Για λόγους συντομίας παρουσιάζουμε εδώ μόνο τις 2 αυτές περιπτώσεις.

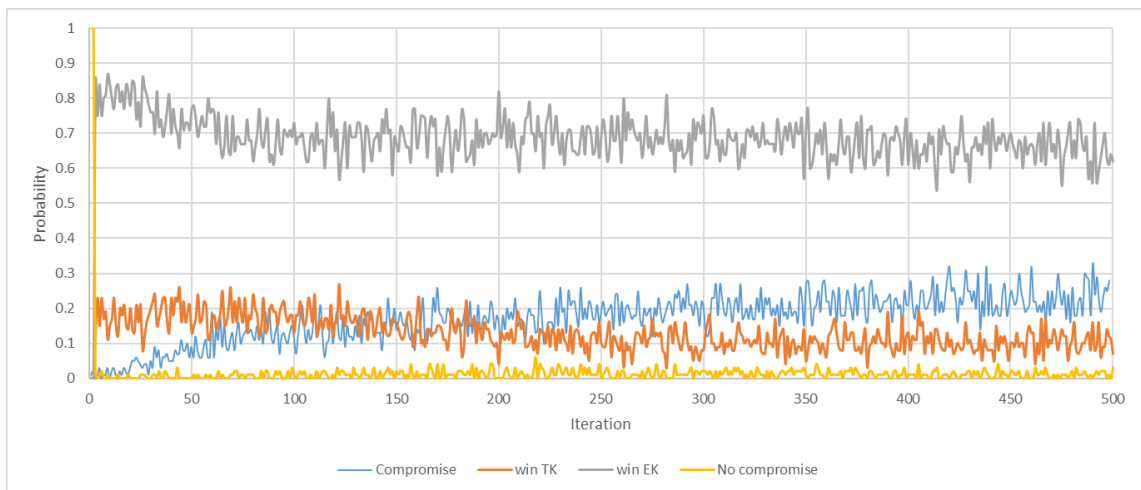
1^η Περίπτωση SARSA

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
EK	0.3	0.5
TK	0.3	0.5

Πίνακας 5.3.1

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.3.1



Σχήμα 5.3.1

Αποτελέσματα προσομοίωσης SARSA με τις παραμέτρους του πίνακα 5.3.1

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας μαθαίνουν αρκετά να συμβιβάζονται και υπάρχει μια αυξητική τάση στην πιθανότητα συμβιβασμού χωρίς όμως να φτάνουμε σε ικανοποιητικά επίπεδα. Ο λόγος πιθανόν να είναι το χαμηλό LR που έχουμε και δεν αφήνει τους πράκτορες να εκπαιδευτούν αρκετά. Η γραφική αυτή δείχνει επίσης πως οι νίκες του EK πράκτορα μειώνονται με την πάροδο των επαναλήψεων

Όπως φαίνεται στο σχήμα 5.3.1 σε σύγκριση με το σχήμα 5.2.1 βλέπουμε πως με το τέλος των επαναλήψεων έχουμε μια πιθανότητα μεγαλύτερη από 10% υψηλότερη για συμβιβασμό. Παράλληλα με την αύξηση της πιθανότητας του συμβιβασμού έχουμε και την μείωση των μονόπλευρων νικών. Αυτή η διαφορά μεταξύ του Q-learning agent και του SARSA agent πιθανόν να οφείλεται στο ότι ο SARSA πράκτορας λαμβάνει υπόψιν του στην ενημέρωση της συνάρτησης αξίας και τις ενέργειες εξερεύνησης οι οποίες είναι δυνατό αρκετές φορές να οδηγούν στον συμβιβασμό και όχι μόνο στους στόχους με προσωπικό συμφέρον.

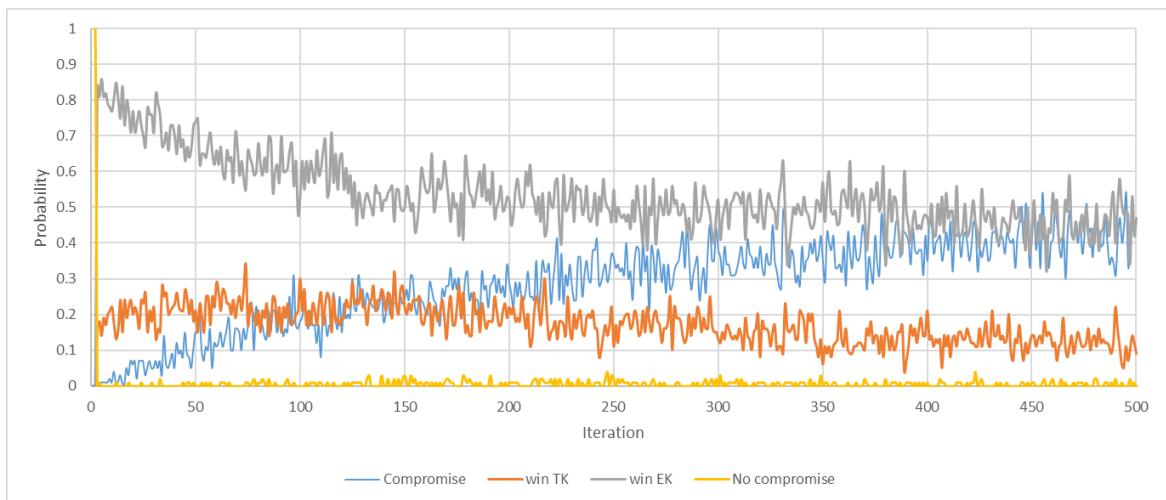
2^η Περίπτωση SARSA

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor
EK	0.7	0.7
TK	0.7	0.7

Πίνακας 5.3.2

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.3.2



Σχήμα 5.3.2

Αποτελέσματα προσομοίωσης SARSA με τις παραμέτρους του πίνακα 5.3.2

Στο σχήμα αυτό βλέπουμε πως με αυτές τις παραμέτρους οι πράκτορες μας εκπαιδεύονται αρκετά και μαθαίνουν να συμβιβάζονται. Το ποσοστό συμβιβασμού συγκλίνει με τις νίκες του EK πράκτορα οι οποίες και μειώνονται αρκετά από την αρχική επανάληψη.

Όπως φαίνεται στο σχήμα 5.3.2 σε σύγκριση με το σχήμα 5.2.7 βλέπουμε πως με το τέλος των επαναλήψεων έχουμε μια πιθανότητα συμβιβασμού κοντά στο 50% στην περίπτωση όπου χρησιμοποιούμε τον SARSA πράκτορα σε αντίθεση με ένα 30% με την χρήση Q-learning agent. Αυτό πιθανόν και πάλι να οφείλετε στον τρόπο που μαθαίνει ο SARSA πράκτορας και αφού λαμβάνει υπόψιν του στις ενημερώσεις και τις

ενέργειες εξερεύνησης. Επίσης η νίκες του ΕΚ πράκτορα μειώνονται αρκετά σε σχέση με τις αρχικές επαναλήψεις.

5.4 Προσθήκη Άδηλων Πηγών Ενίσχυσης

Για την εξέταση της επιρροής των ΑΠΕ θα χρησιμοποιήσουμε μόνο τον Q-learning πράκτορα αφού έχουν παρόμοια αποτελέσματα με τον SARSA. Οι διάφοροι εξωτερικοί παράγοντες υλοποιήθηκαν ως μέρος της κλάσης ΑΠΕ, ο κάθε πράκτορας μετά από μια πράξη του καλεί την δημόσια συνάρτηση της κλάσης αυτής δίνοντας της την τρέχων κατάσταση, την προσφορά, και την απάντηση που έχει δώσει ο πράκτορας. Η συνάρτηση αυτή με την σειρά της δίνει στον κάθε εξωτερικό παράγοντα (υλοποιούνται εσωτερικά της κλάσης) τα στοιχεία αυτά και επιστρέφει ο καθένας το δικό του Ενισχυτικό Σήμα. Έπειτα, η συνάρτηση υπολογίζει το συνολικό σήμα και το επιστρέφει στον πράκτορα ως ένα σήμα.

Πιο κάτω θα παρουσιάσουμε 2 περιπτώσεις με τα σημαντικότερα αποτελέσματα με την χρήση των ΑΠΕ.

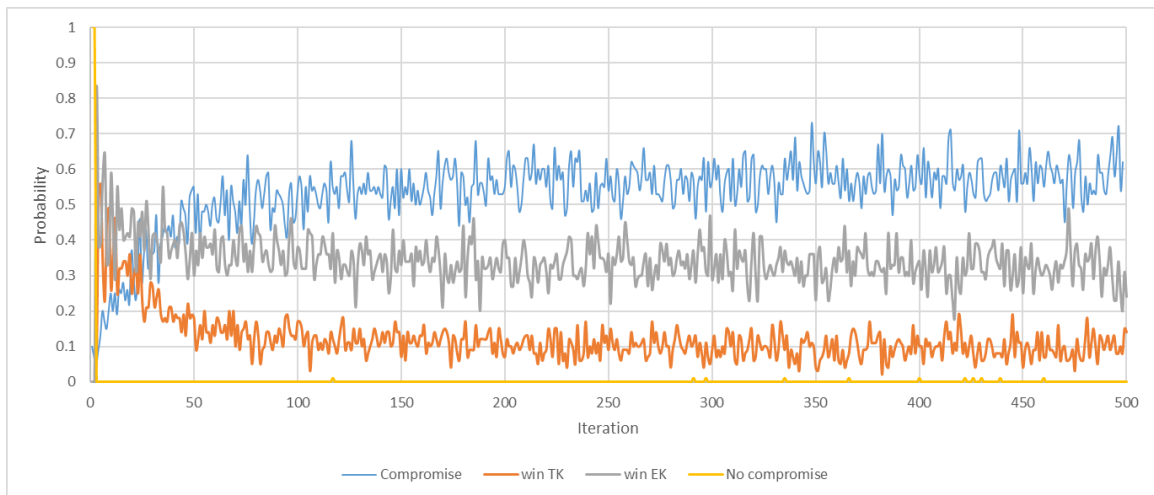
1^η Περίπτωση ΑΠΕ

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor	External Weight
EK	0.5	0.5	1
TK	0.5	0.5	1

Πίνακας 5.4.1

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.4.1



Σχήμα 5.4.1

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.4.1

Στο σχήμα αυτό βλέπουμε πως η χρήση των εξωτερικών παραγόντων σε συνδυασμό με αυτές τις παραμέτρους συμβάλει κατά πολύ στην αύξηση του ποσοστού συμβιβασμού. Το ποσοστό συμβιβασμού ξεπερνάει τις νίκες και των 2 πρακτόρων φτάνοντας περίπου στο 70%.

Στην πρώτη περίπτωση χρησιμοποιήσαμε τις μεταβλητές με τις οποίες ο Q-learning είχε ένα από τα καλύτερα αποτελέσματα όσο αφορά το ποσοστό συμβιβασμού. Με την χρήση των ΑΠΕ βλέπουμε πως οι νίκες του ΕΚ πράκτορα μειώνονται και παράλληλα η πιθανότητα για συμβιβασμό αυξάνεται κατά πολύ φτάνοντας μέχρι και στο 70% (σχήμα 5.4.1). Φαίνεται πως η συμβολή των εξωτερικών παραγόντων επηρεάζει αρκετά τους 2 εμπλεκόμενους προς τον συμβιβασμό.

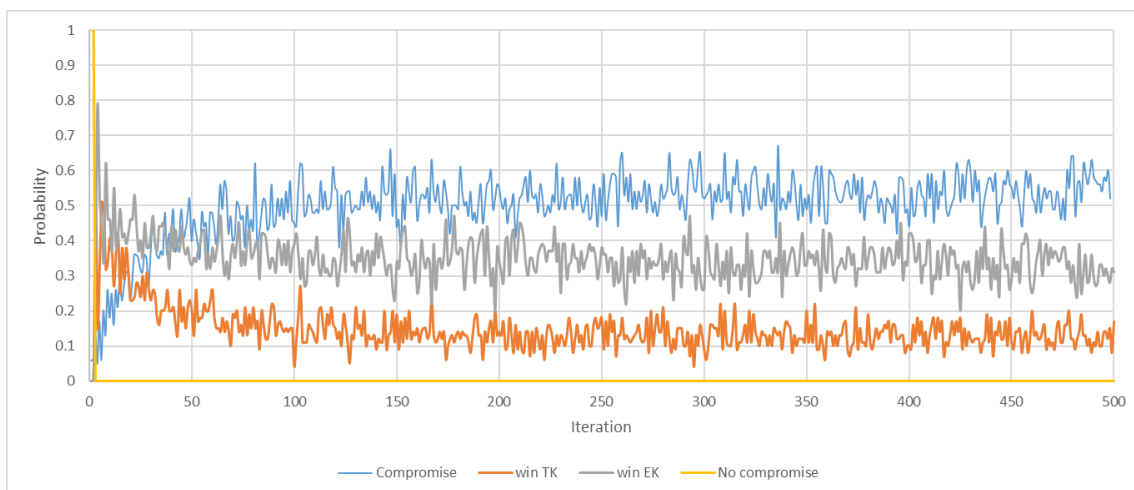
2^η Περίπτωση ΑΠΕ

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor	External Weight
ΕΚ	0.8	0.5	1
ΤΚ	0.8	0.5	1

Πίνακας 5.4.2

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.4.2



Σχήμα 5.4.2

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.4.2

Στο σχήμα αυτό βλέπουμε πως η χρήση των εξωτερικών παραγόντων σε συνδυασμό με αυτές τις παραμέτρους συμβάλει στην αύξηση του ποσοστού συμβιβασμού. Η αύξηση όμως του LR φαίνεται να επηρέασε την αύξηση του ποσοστού συμβιβασμού και οι πράκτορες και πάλι φαίνεται να βλέπουν αρκετά το προσωπικό τους συμφέρον χωρίς ωστόσο να τους «επιτρέπουν» οι ΑΠΕ να ανεβάσουν ιδιαίτερα τα ποσοστά μονόπλευρης νίκης.

Ένας άλλος συνδυασμός παραμέτρων που είχε ως αποτέλεσμα σχετικά ψηλά ποσοστά συνεργασίας ήταν αυτός που φαίνεται στον πίνακα 5.2.4. και σε αυτή την περίπτωση βλέπουμε να αυξάνονται τα ποσοστά συμβιβασμού από ένα σχετικά αρχικό σημείο, από την 50^η επανάληψη έχουμε ήδη περίπου 45% ποσοστό συμβιβασμού, φτάνοντας στις 500 επαναλήψεις με περίπου 60% πιθανότητα συμβιβασμού.

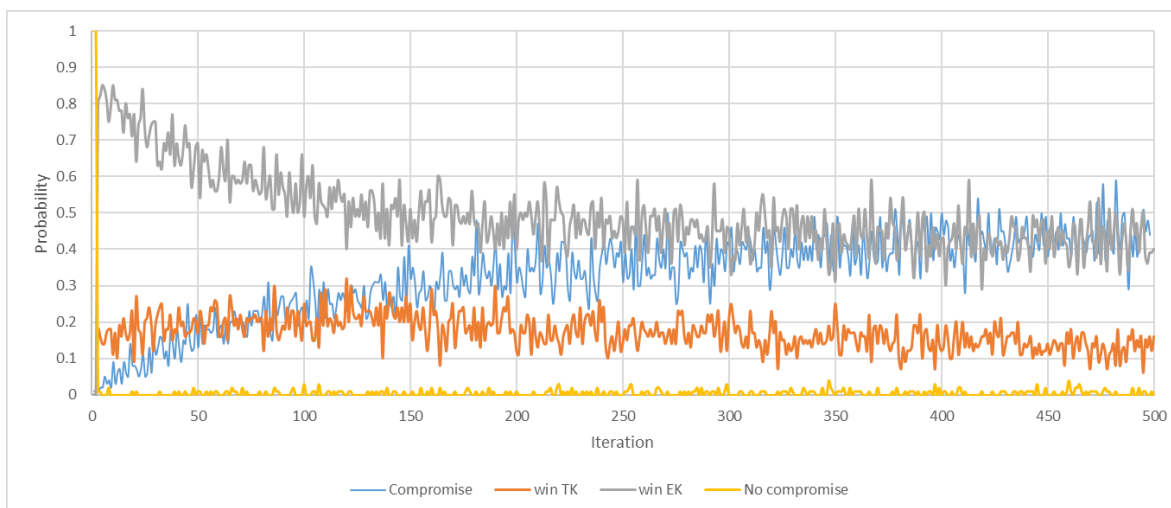
3^η Περίπτωση ΑΠΕ

Παράμετρον/ Πράκτορας	Learning Rate	Discount Factor	External Weight
EK	0.3	0.5	1
TK	0.3	0.5	1

Πίνακας 5.4.3

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.4.3



Σχήμα 5.4.3

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρου του πίνακα 5.4.3

Στο σχήμα αυτό βλέπουμε πως η χρήση των εξωτερικών παραγόντων σε συνδυασμό με αυτές τις παραμέτρους καταλήγει στην σύγκλιση και πάλι του ποσοστού συμβιβασμών με το ποσοστό των νικών του ΕΚ. Το πολύ χαμηλό LR σε αυτή την περίπτωση φαίνεται πως επηρεάζει α αποτελέσματα αυτά αφού πιθανόν οι πράκτορες να μην μαθαίνουν και τόσο από τις προσωπικές τους εμπειρίες αλλά βασίζονται κυρίως στο ενισχυτικό σήμα των ΑΠΕ.

Σε αυτή την περίπτωση θέλαμε να συγκρίνουμε αποτελέσματα χωρίς την χρήση ΑΠΕ τα οποία δεν είχαν πολύ ψηλά ποσοστά συμβιβασμού με αποτελέσματα που κάνουν χρήση των ΑΠΕ. Επιλέξαμε να παρουσιάσουμε την γραφική με της μεταβλητές που φαίνονται στον πίνακα 5.2.1 τα οποία είχαν πολύ χαμηλό ποσοστό συμβιβασμού. Όπως φαίνεται ξεκάθαρα στο σχήμα 5.4.4 με την χρήση ΑΠΕ έχουμε πολύ καλύτερα αποτελέσματα. Αυτό πιθανόν να οφείλεται στο ότι οι πράκτορες ακόμη και με χαμηλό learning rate επηρεάζονται από το ενισχυτικό σήμα των ΑΠΕ και οδηγούνται προς τον συμβιβασμό.

5.5 Προσθήκη Επιπρόσθετης Αμοιβής

Στην συνέχεια θέλαμε να δοκιμάσουμε κατά πόσο η επιβράβευση ή η τιμωρία σε όλες τις ενέργειες που ακολουθήθηκαν μέχρι να φτάσουμε στο τελικό αποτέλεσμα θα μπορούσαν να βοηθήσουν να έχουμε μεγαλύτερη πιθανότητα συμβιβασμού.

Στον πίνακα 5.5.1 και στο σχήμα 5.5.2 έχουμε την περίπτωση κατά την οποία οι πράκτορες όταν φτάσουν σε συμβιβασμό όλες τους οι διαδοχικές ενέργειες που έκαναν μέχρι τότε επιβραβεύονται με ένα ενισχυτικό σήμα με βάρος 3 και όταν φτάσουν σε μη συμβιβασμό όλες τους οι διαδοχικές ενέργειες που έκαναν μέχρι τότε τιμωρούνται με ένα ενισχυτικό σήμα με βάρος -3.

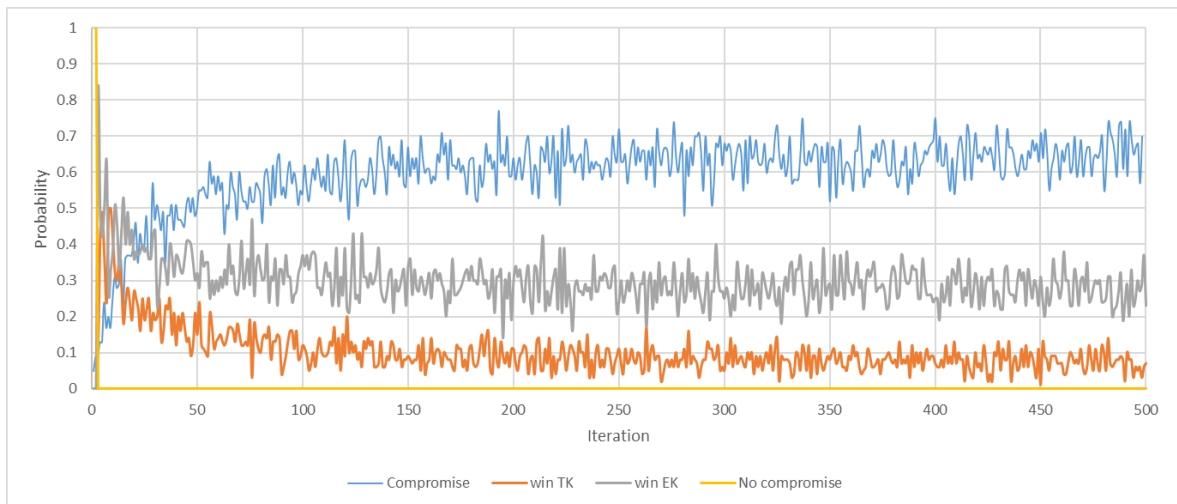
Δοκιμάσαμε τις παραμέτρους που φαίνονται στον πίνακα 5.4.1. όπου και είχαμε ένα από τα πιο ψηλά ποσοστά συμβιβασμού. Αρχικά, ενεργοποιήσαμε την επιλογή όπου προσφέρει αμοιβή ή τιμωρία σε όλες τις διαδοχικές ενέργειες όταν καταλήξουμε σε συμβιβασμό ή σε μη συμβιβασμό. Τα αποτελέσματα αυτής της δοκιμής φαίνονται να είναι ελαφρώς καλύτερα όπως δείχνει και το σχήμα 5.5.1, ξεπερνώντας μερικές φορές και το 70% για πιθανότητα συμβιβασμού.

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor	External Weight	Compromise or no- compromise reward	One side win
EK	0.5	0.5	1	True	False
TK	0.5	0.5	1	True	False

Πίνακας 5.5.1

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.5.1



Σχήμα 5.5.1

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.5.1

Στο σχήμα αυτό βλέπουμε πως η χρήση της επιπρόσθετης αμοιβής στις περιπτώσεις που η επανάληψη καταλήξει σε συμβιβασμό ή μη συμβιβασμό, συμβάλει ακόμη περισσότερο στην αύξηση του ποσοστού συμβιβασμού το οποίο σε μερικές περιπτώσεις ξεπερνάει το 70%. Οι μονόπλευρες νίκες και των δύο πρακτόρων μειώνονται σημαντικά.

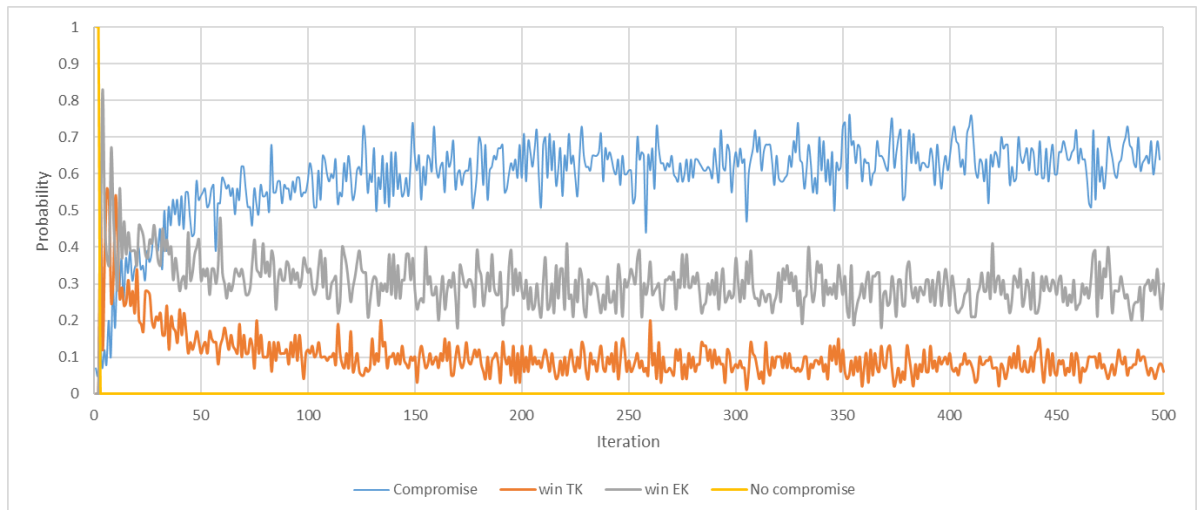
Έπειτα, δοκιμάζοντας να ενεργοποιήσουμε και την επιλογή όπου οι διαδοχικές ενέργειες ενός πράκτορα ο οποίος κατέληξε σε νίκη τιμωρούνται, παρατηρήσαμε πως δεν υπήρξε κάποια σημαντική διαφορά. Όλες οι ενέργειες του πράκτορα που τον οδήγησαν σε δική του νίκη, δηλαδή όχι ικανοποίηση του αντιπάλου, τιμωρούνται με Ενισχυτικό σήμα -1. Αυτό δεν φαίνεται να επηρεάζει ιδιαίτερα τον πράκτορα στην απόφαση που παίρνει.

Παράμετροι/ Πράκτορας	Learning Rate	Discount Factor	External Weight	Compromise or no- compromise reward	One side win
EK	0.5	0.5	1	True	True
TK	0.5	0.5	1	True	True

Πίνακας 5.5.2

Πίνακας παραμέτρων

Ο πίνακας παρουσιάζει τις παραμέτρους που χρησιμοποιήθηκαν για τα αποτελέσματα που φαίνονται στο σχήμα 5.5.2



Σχήμα 5.5.2

Αποτελέσματα προσομοίωσης Q-learning με τις παραμέτρους του πίνακα 5.5.2

Στο σχήμα αυτό βλέπουμε πως η χρήση της επιπρόσθετης αμοιβής ή τιμωρίας τόσο όταν η επανάληψη καταλήξει σε συμβιβασμό ή μη συμβιβασμό αντίστοιχα αλλά και τιμωρίας όταν η επανάληψη καταλήξει σε μονόπλευρη νίκη. Φαίνεται πως αυτή η επιλογή δεν επηρεάζει το αποτέλεσμα σε σχέση με την χρήση των υπόλοιπων επιλογών και έχουμε παρόμοια αποτελέσματα και χωρίς την χρήση της.

Κεφάλαιο 6

Συμπεράσματα και μελλοντική εργασία

6.1 Συμπεράσματα	71
6.2 Μελλοντική Εργασία	73

6.1 Συμπεράσματα

Στόχος της εργασίας αυτής ήταν η προσομοίωση του Κυπριακού προβλήματος, ενός πολύπλοκου προβλήματος, με την χρήση EM και του Bargaining game. Ασχοληθήκαμε με την εδαφική πτυχή του θέματος η οποία είχε μερικούς ξεκάθαρους αριθμούς έτσι ώστε να μπορούμε να μετρούμε τα αποτελέσματα καλύτερα. Θέλαμε να δούμε κατά πόσο οι πράκτορες μαθαίνουν να συμβιβάζονται σε ένα πρόβλημα στο οποίο στην πραγματικότητα ακόμη δεν τα έχουν καταφέρει.

Έπειτα, ένας δεύτερος στόχος ήταν η εισαγωγή των εξωτερικών παραγόντων που επηρεάζουν την έκβαση των διαπραγματεύσεων, όχι ως άμεσοι παίκτες αλλά ως έμμεσοι. Έτσι αποφασίσαμε να τους εντάξουμε αξιοποιώντας κάποιες γνώσεις από την ψυχολογία και χρησιμοποιώντας τους ως Άδηλες Πηγές Ενίσχυσης.

Με την δοκιμή του συστήματος μας σε μια πρώτη φάση και δοκιμάζοντας διάφορες παραμέτρους βλέπουμε πως η αντιστοίχιση του προβλήματος στο σύστημα ήταν εφικτή και είχε λογικά αποτελέσματα. Παρατηρήσαμε πως στα αρχικά στάδια είχαμε διαφωνίες οπότε και κατέληγαν σε μη συμβιβασμό ή πολύ συχνά σε μονόπλευρη νίκη. Όμως με την πάροδο αρκετών εποχών και την επιλογή των κατάλληλων παραμέτρων βλέπαμε πως οι πράκτορες μάθαιναν να συμβιβάζονται, όχι 100% κάτι το οποίο δεν θα ήταν ρεαλιστικό με βάση το πρόβλημα μας αλλά είχαν αρκετά υψηλά ποσοστά συμβιβασμού.

Επιπρόσθετα, παρατηρήσαμε και την σημαντικότητα της σωστής επιλογής των παραμέτρων στην εξέλιξη και στα αποτελέσματα του συστήματος. Μερικές φορές όταν ο πράκτορας βλέπει μακροπρόθεσμα δεν είναι τόσο θετικό για το κοινό καλό. Όπως επίσης και πολύ χαμηλοί ρυθμοί μάθησης δεν επέτρεπαν συνήθως στον πράκτορα να εκπαιδευτεί επαρκώς με αποτέλεσμα να μην βελτιώνονται οι επιλογές του. Όλες οι δοκιμές αυτές φαίνονται αναλυτικά στα υποκεφάλαια 5.2 και 5.3 . Στο καθένα από αυτά κεφάλαια δοκιμάσαμε διαφορετικές παραμέτρους με διαφορετικό πράκτορα.

Αυτό που μπορούμε να εξάγουμε από τα περισσότερα αποτελέσματα στα δύο αυτά υποκεφάλαια είναι το πως συνήθως ο ΕΚ πράκτορας είχε περισσότερες νίκες. Αυτό οφείλετε στο γεγονός ότι απαιτεί πολύ λιγότερες μονάδες εδάφους από τον ΤΚ πράκτορα για να είναι ικανοποιημένος. Έτσι ενώ ο ΤΚ πράκτορας μετά την 94^η επανάληψη, είναι αδύνατο να είναι ικανοποιημένος , ο ΕΚ πράκτορας συνεχίζει μέχρι και την 296^η επανάληψη να διεκδικεί την δική του νίκη την οποία και παίρνει. Είναι ξεκάθαρο όμως πως τις περισσότερες φορές ο ΤΚ πράκτορας μαθαίνει να μην αφήνει τόσες φορές τον ΕΚ να είναι ικανοποιημένος μόνο αυτός οπότε και χαμηλώνει το ποσοστό των νικών του ΕΚ.

Ακολουθώντας, με την εισαγωγή των εξωτερικών παραγόντων, από τους οποίους οι περισσότεροι επιθυμούν μια λύση συμβιβασμού, ως ΑΠΕ, τα αποτελέσματα αλλάζουν. Βλέπουμε το ποσοστό των νικών του ΕΚ να μειώνεται ακόμη περισσότερο το οποίο πιθανόν να οφείλεται στις τιμωρίες/κυρώσεις που δέχεται από τις χώρες αυτές και παράλληλα έχουμε περαιτέρω αύξηση στην πιθανότητα συμβιβασμού. Είναι ξεκάθαρο ότι ο ΕΚ πράκτορας πρέπει να μάθει να μην βλέπει μόνο το προσωπικό του συμφέρον, διότι στο σύστημα μας με βάση τις απαιτήσεις του έχει πλεονέκτημα, αλλά να επιδιώκει το κοινό καλό. Το ίδιο ισχύει και για τον ΤΚ πράκτορα ο οποίος όμως λόγω των υψηλών απαιτήσεων δεν είχε τόσο υψηλά ποσοστά νίκης. Αυτό φαίνεται να επιτυγχάνεται μέσα από την χρήση των ΑΠΕ, ασχέτως αν οι 2 από τις χώρες που χρησιμοποιούνται ενθαρρύνουν την νίκη του πράκτορα όπου έχουν συμφέρον.

Οι ΑΠΕ τιμωρούν και επιβραβεύουν τους δύο πράκτορες αναλόγως με τις ενέργειες τους από την αρχή και αυτό φαίνεται να προσφέρει υψηλά ποσοστά συμβιβασμού σε αρκετά πιο αρχικές επαναλήψεις εκπαίδευσης. Διότι, στις περισσότερες περιπτώσεις που χρησιμοποιούμε τις ΑΠΕ το ποσοστό συμβιβασμού ανεβαίνει ραγδαία στα αρχικά στάδια. Κάτι το οποίο δεν ισχύει και για τις εκπαιδεύσεις χωρίς ΑΠΕ.

Ακόμη, η δοκιμή της καθυστερημένης τιμωρίας/ αμοιβής ήταν κάτι το οποίο είχε καθαρά πειραματικό χαρακτήρα και όμως φαίνεται να επηρεάζει το αποτέλεσμα, έστω και κατά ένα μικρό ποσοστό, και να συμβάλει και αυτή η καθυστερημένη αμοιβή στα υψηλότερα ποσοστά συμβιβασμού των δύο πρακτόρων. Με την χρήση αυτής της επιλογής πετύχαμε το υψηλότερο ποσοστό συμβιβασμού, από όλες τις προσπάθειες με διαφορετικές παραμέτρους, το οποίο ήταν λίγο πάνω από 70%.

Τέλος, το σύστημα αυτό θα μπορούσε να αποτελέσει ένα σημαντικό βοήθημα για την διαπραγματευτική ομάδα αφού χρησιμοποιώντας το σύστημα μπορούμε να έχουμε προσομοίωση εκατοντάδων διαφορετικών περιπτώσεων και να δούμε ποιες από αυτές έχουν τα επιθυμητά αποτελέσματα χωρίς να καθυστερούμε ή να έχουμε κακές εξελίξεις στην πραγματικότητα. Έτσι, μπορούμε να καταλάβουμε και ποιες επιλογές κάνουν τον αντίπαλο πράκτορα να γίνεται περισσότερο εγωιστής και να αρνείται οποιαδήποτε προσφορά με σκοπό να τις αποφύγουμε. Υπάρχει η δυνατότητα να αναλύσουμε τα ενισχυτικά σήματα των ΑΠΕ και να τα χρησιμοποιήσουμε έτσι ώστε να γνωρίζουν και οι εξωτερικοί παράγοντες με βάση ποιες αποφάσεις θα πρέπει να τιμωρούν ή να επιβραβεύουν την μια πλευρά για να έχουμε καλύτερα αποτελέσματα.

6.2 Μελλοντική Εργασία

Η εργασία αυτή έχει την δυνατότητα να επεκταθεί με διάφορους τρόπους και σε διαφορετικά σημεία. Λόγω του ότι είναι μια από τις λίγες εργασίες σε τέτοιο θέμα με την χρήση bargaining game χρειάζονται αρκετές δοκιμές για να διαφανεί τι λειτουργεί καλύτερα σε αυτή την περίπτωση.

Αρχικά, θα μπορούσαν να δοκιμαστούν διαφορετικοί αλγόριθμοι μάθησης για τους πράκτορες μας. Υπάρχουν αρκετές μέθοδοι ενισχυτικής μάθησης οι οποίοι δεν είναι ξεκάθαρο το πώς θα δρούσαν σε ένα τέτοιο πρόβλημα οπότεν πρέπει να δοκιμαστούν για να εξαχθούν αποτελέσματα.

Επίσης, θα μπορούσαμε να δοκιμάσουμε διαφορετικές τιμές για το ϵ της e -greedy μεθόδου αλλά και να δοκιμαστούν άλλες τακτικές εκτός από την e -greedy για την ισορροπία μεταξύ την εξερεύνησης και αξιοποίησης των πρακτόρων.

Ακόμη, υπάρχει η δυνατότητα να προστεθεί μια εξήγηση για τον κάθε γύρω και που ουσιαστικά κρίθηκε το αποτέλεσμα αυτού του γύρου μέσω κάποιων στατιστικών από όλες τις αποφάσεις.

Τέλος, θα ήταν δυνατό να χρησιμοποιηθεί ένα άλλο παίγνιο αντικαθιστώντας το bargaining game. Αυτό μπορεί να γίνει εύκολα στο σύστημα μας αλλάζοντας απλώς την κλάση bargaining game με το παίγνιο που επιθυμούμε.

Αναφορές

Bellman R.E. (1957) . Dynamic Programming. Princeton University Press, Princeton, New Jersey.

Deci, E. L. (1972). Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of personality and social psychology*, 22(1), 113-120.

Fudenberg, D., & Tirole, J. (1991). *Game theory*. MIT press.

Gaye Delphine Banfield (2006), *Simulation of Self-Control through Precommitment Behaviour in an Evolutionary System*, PhD thesis, School of Computer Science and Information Systems, Birkbeck College, University of London

Hu, J., & Wellman, M. P. (1998, June). Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML* (Vol. 98, pp. 242-250).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4, 237-285

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994* (pp. 157-163). Morgan Kaufmann.

Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247), 335-341.

Nash Jr, J. F. (1950). Equilibrium points in N-person games. *Proceedings of the national academy of sciences*, 36(1), 48-49.

Nowé, A., Vrancx, P., & Hauwere, Y. M. D. (2012). Game theory and multi-agent reinforcement learning. In *Reinforcement Learning* (pp. 441-470). Springer, Berlin, Heidelberg.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, 50(1), 97-109.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1), 54-67.

Tesauro, G. (1991). Practical issues in temporal difference learning. In *Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS'91)*. Morgan Kaufmann Publishers Inc., 259–266.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton. Princeton University Press, 1947, 1953.

Watkins, C.J. and Dayan, P., 1992. Q-learning. *Machine learning*, 8(3), pp.279-292.

Σχέδιο Αναν 5,

<https://www.pio.gov.cy/assets/pdf/cyproblem/%20%CE%91%CE%BD%CE%AC%CE%BD%205.pdf>

Ανδρέας Δ. Συμεού (2020). Κυπριακό ώρα μηδέν : το αγκάθι του περιουσιακού , Παπαζήση

Βασιλειάδης Β. (2007) Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων όπως εφαρμόζεται σε Παίγνια Γενικού Αθροίσματος. Ατομική Διπλωματική Εργασία, Τμήμα Πληροφορικής, Πανεπιστήμιο Κύπρου.

Γεωργίου Σ. (2007). Του παιδιού και του σχολείου, Ατραπός, Κεφάλαιο 6, σελίδες 209-244.

Ιωάννης Παρτάλας (2009), Μέθοδοι Ενισχυτικής Μάθησης σε Συστήματα

Πρακτόρων, Διδακτορική Διατριβή, Τμήμα Πληροφορικής, Αριστοτέλειο Πανεπιστήμιο
Θεσσαλονίκης

Καραολής, Γ. (2010). Προσέγγιση του κυπριακού προβλήματος με ενισχυτική μάθηση
πολλαπλών πρακτόρων σε παίγνια γενικού αθροίσματος. Διπλωματική Εργασία, Τμήμα
Πληροφορικής, Πανεπιστήμιο Κύπρου.

Πολύβιος Γ. Πολυβίου (2020). Το κυπριακό πρόβλημα : παρελθόν, παρόν και μέλλον,
Επίκεντρο

Παράρτημα Α

Οδηγίες χρήσης κώδικα.

Ο κώδικας γράφτηκε και δοκιμάστηκε με Java 14 (SDK 14.0.1).

Για να δουλέψει ο κώδικας **απαιτεί** το αρχείο parameters.txt όπως φαίνεται στο παράδειγμα πιο κάτω.

Parameters.txt:

```
iterations = 500  
area = 362  
satisfied ek = 66  
satisfied tk = 278  
q learning  
learning rate tk = 0.5  
learning rate ek = 0.5  
discount factor tk = 0.5  
discount factor ek = 0.5  
externals lr tk = 1  
externals lr ek = 1
```

Για την χρήση SARSA πρακτόρων αλλάζουμε το ‘q learning’ με ‘sarsa’. Η σειρά των παραμέτρων που εμφανίζονται στο αρχείο δεν έχει σημασία.

Εκτέλεση προγράμματος

Όταν έχουμε όλα τα αρχεία (8, φαίνονται πιο κάτω στο Παράρτημα Β) σε ένα package προσθέτουμε και το parameters.txt και έπειτα τρέχουμε την κλάση BargainingGame.java η οποία περιέχει την main function.

Έξοδος προγράμματος

Τα αποτελέσματα τυπώνονται τόσο στην κονσόλα όσο και σε ένα αρχείο με το όνομα outputXX.txt όπου XX μερικές από τις παραμέτρους που εισαγάγαμε στο parameters.txt (sarsa, lr_ek, discount_f_ek, external_lr_ek, external_prev_actions).

Παράρτημα Β

Source Code

BargaininigGame.java

```
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;
import java.util.ArrayList;
import java.util.Locale;
import java.util.Scanner;

public class BargainingGame {
    private static Scanner scan;

    private static int iterations=0;
    private static int area = 0;
    private static int satis_tk = 0;
    private static int satis_ek = 0;
    private static boolean sarsa=false;
    private static double lr_ek=0;
    private static double lr_tk=0;
    private static double external_lr_tk=0;
    private static double external_lr_ek=0;
    private static double discount_f_tk=0;
    private static double discount_f_ek=0;
    private static boolean external_prev_actions= false;
    private static boolean external_prev_wins= false;

    private static void readParameters(String fname){
        File file= new File(fname);
        try {
            scan = new Scanner(file);
        } catch (IOException e){
```

```

        System.out.println("Parameters file problem, aborting");
        System.exit(0);
    }

    while(scan.hasNextLine()){
        String line = scan.nextLine();
        String splited[]=line.split(" ");
        line = line.toLowerCase(Locale.ROOT);

        if (line.contains("iterations"))
            iterations = Integer.parseInt(splited[splited.length-1]);
        if (line.contains("area"))
            area = Integer.parseInt(splited[splited.length-1]);
        if (line.contains("satisfied ek"))
            satis_ek = Integer.parseInt(splited[splited.length-1]);
        if (line.contains("satisfied tk"))
            satis_tk = Integer.parseInt(splited[splited.length-1]);
        if (line.contains("sarsa"))
            sarsa = true;
        if (line.contains("q learning"))
            sarsa = false;
        if (line.contains("learning rate tk"))
            lr_tk = Double.parseDouble(splited[splited.length-1]);
        if (line.contains("learning rate ek"))
            lr_ek = Double.parseDouble(splited[splited.length-1]);
        if (line.contains("externals lr tk"))
            external_lr_tk = Double.parseDouble(splited[splited.length-1]);
        if (line.contains("externals lr ek"))
            external_lr_ek = Double.parseDouble(splited[splited.length-1]);
        if (line.contains("discount factor tk"))
            discount_f_tk = Double.parseDouble(splited[splited.length-1]);
        if (line.contains("discount factor ek"))
            discount_f_ek = Double.parseDouble(splited[splited.length-1]);
    }

```

```

}

public static void main(String[] args) {

    readParameters("parameters.txt");

    Agent tk;
    Agent ek;
    if (!sarsa) {
        tk = new QLearningAgent('t', area, satis_tk, lr_tk, discount_f_tk);
        ek = new QLearningAgent('e', area, satis_ek, lr_ek, discount_f_ek);
    } else {
        tk = new SARSAAgent('t', area, satis_tk, lr_tk, discount_f_tk);
        ek = new SARSAAgent('e', area, satis_ek, lr_ek, discount_f_ek);
    }

    int comp = 0;
    int winek = 0;
    int wintk = 0;
    boolean debug = false;
    int curr_area;
    int turninitial = (int) (Math.random() * 2);
    int probability[][] = new int[iterations][4];
    for (int i = 0; i < iterations; i++)
        for (int j = 0; j < 4; j++)
            probability[i][j] = 0;

    int tries = 100;
    for (int w = 0; w < tries; w++) {
        for (int i = 0; i < iterations; i++) {
            curr_area = area;
            turninitial = (++turninitial) % 2;
            int turn = turninitial;
            int decision = 0;
            int off;

```

```

double rew;
Environment env = new Environment();
Action act;
ArrayList<Action> ekactions = new ArrayList<>();
ArrayList<Action> tkactions = new ArrayList<>();
while (curr_area > 0) {

    if (turn == 0) {
        if (curr_area == area)
            off = ek.make_offer(curr_area);
        else
            off = decision;

        if (debug) System.out.print("Offer from ek: " + off);

        act = tk.decide_for_offer(off, curr_area);
        tkactions.add(act);
        if (act.decision.equals("Accept")) {
            decision = -1;
        } else {
            decision = tk.make_offer(curr_area - 1);
        }
        rew = env.reward(off, decision, curr_area, tk);
        rew += external_lr_tk * ExternalFactors.calculateReward(curr_area, off, decision,
tk, ek);

        double test = tk.updatevalue(act, rew);

        act.setValue_function(test);
        if (debug) System.out.println(" Decision of tk: " + decision);
    } else {
        if (curr_area == area)
            off = tk.make_offer(curr_area);
        else
            off = decision;
    }
}

```

```

    if (debug) System.out.print("Offer from tk: " + off);
    act = ek.decide_for_offer(off, curr_area);
    ekactions.add(act);
    if (act.decision.equals("Accept")) {
        decision = -1;
    } else {
        decision = ek.make_offer(curr_area - 1);
    }
    rew = env.reward(off, decision, curr_area, ek);
    rew += external_lr_ek * ExternalFactors.calculateReward(curr_area, off, decision,
ek, tk);

    double test = ek.updatevalue(act, rew);

    act.setValue_function(test);
    if (debug) System.out.println(" Decision of ek: " + decision);
}
ek.curr_state++;
tk.curr_state++;
if (decision == -1) {
    if (turn == 0) {
        if (off >= satis_tk && (curr_area - off) >= satis_ek) {
            if (external_prev_actions){
                for (int z = 0; z < ekactions.size(); z++)
                    ekactions.get(z).addReward(external_lr_ek * 3);
                for (int z = 0; z < tkactions.size(); z++)
                    tkactions.get(z).addReward(external_lr_tk * 3);
            }
            System.out.println("compromise");
            comp++;
            probability[i][0]++;
        } else if (off >= satis_tk && (curr_area - off) < satis_ek) {
            System.out.println("Win TK available area = " + curr_area + " tk area: " + off);
            wintk++;
            if (external_prev_wins)
                for (int z = 0; z < tkactions.size(); z++)

```

```

        tkactions.get(z).addReward(external_lr_tk * -1);
        probability[i][1]++;
    } else if ((curr_area - off) >= satis_ek) {
        System.out.println("Win EK available area = " + curr_area + " ek area: " +
(curr_area - off));
        winek++;
        if (external_prev_wins)
            for (int z = 0; z < ekactions.size(); z++)
                ekactions.get(z).addReward(external_lr_ek * -1);
        probability[i][2]++;
    } else {
        System.out.println(i + " No Compromise offer tk: " + off + " ek: " + (curr_area
- off));

        if (external_prev_actions) {
            for (int z = 0; z < ekactions.size(); z++)
                ekactions.get(z).addReward(external_lr_ek * -3);
            for (int z = 0; z < tkactions.size(); z++)
                tkactions.get(z).addReward(external_lr_tk * -3);
        }
        probability[i][3]++;
    }
} else {
    if (off >= satis_ek && (curr_area - off) >= satis_tk) {
        System.out.println("compromise");
        if (external_prev_actions) {
            for (int z = 0; z < ekactions.size(); z++)
                ekactions.get(z).addReward(external_lr_ek * 3);
            for (int z = 0; z < tkactions.size(); z++)
                tkactions.get(z).addReward(external_lr_tk * 3);
        }
        comp++;
        probability[i][0]++;
    } else if (off >= satis_ek && (curr_area - off) < satis_tk) {
        System.out.println("Win EK available area = " + curr_area + " ek area: " + off);
        winek++;
        if (external_prev_wins)

```

```

        for (int z = 0; z < ekactions.size(); z++)
            ekactions.get(z).addReward(external_lr_ek * -1);
        probability[i][2]++;
    } else if ((curr_area - off) >= satis_tk) {
        System.out.println("Win TK available area = " + curr_area + " tk area: " +
(curr_area - off));
        wintk++;
        if (external_prev_wins)
            for (int z = 0; z < tkactions.size(); z++)
                tkactions.get(z).addReward(external_lr_tk * -1);
            probability[i][1]++;
    } else {
        System.out.println(i + " No Compromise offer ek: " + off + " tk: " + (curr_area
- off));

        if (external_prev_actions) {
            for (int z = 0; z < ekactions.size(); z++)
                ekactions.get(z).addReward(external_lr_ek * -3);
            for (int z = 0; z < tkactions.size(); z++)
                tkactions.get(z).addReward(external_lr_tk * -3);
        }
        probability[i][3]++;
    }
}
break;
} else if (curr_area == 1) {
    System.out.println("No compromise");
    if (external_prev_actions) {
        for (int z = 0; z < ekactions.size(); z++)
            ekactions.get(z).addReward(external_lr_ek * -3);
        for (int z = 0; z < tkactions.size(); z++)
            tkactions.get(z).addReward(external_lr_tk * -3);
    }
    probability[i][3]++;
}
turn = (++turn) % 2;
curr_area--;

```



```

        myWriter.close();
    }
    catch (IOException e) {
        System.out.println("An error occurred while creating the file.");
        e.printStackTrace();
    }
}
}

```

Agent.java

```

import java.util.ArrayList;
import java.util.LinkedList;

public abstract class Agent {

    int curr_state=0;
    double satisfied;
    int init_area;
    double lr;
    double disc_factor;
    char agent_name;
    State stateDecision[];

    abstract int make_offer(int area);
    abstract Action decide_for_offer(int offer, int av_area);
    abstract void reset_state();
    abstract void reset_all();
    abstract int maxQ(boolean next);
    abstract void debugstate();
    abstract double updatevalue(Action act,double reward);
}

```

```
public class SARSAAgent extends Agent{

    Action nextActionG;
    Action nextActionL;
    Action nextActionE;

    public SARSAAgent(char name, int area, double satisfied, double learning_rate, double
discount_f) {
        nextActionG = null;
        nextActionL = null;
        nextActionE = null;
        init_area = area;
        this.satisfied=satisfied;
        lr=learning_rate;
        disc_factor=discount_f;
        stateDecision = new State[area];
        agent_name = name;
        for(int i=0;i<area;i++){
            stateDecision[i]=new State();
            stateDecision[i].actions.add(new Action("Refuse","<0.5",0));
            stateDecision[i].actions.add(new Action("Refuse",">=0.5",0));
            stateDecision[i].actions.add(new Action("Accept","<0.5",0));
            stateDecision[i].actions.add(new Action("Accept",">=0.5",0));

        }

    }

    int maxQO(boolean next,String sat){

        int state = curr_state;
        if (next)
            state=curr_state+1;
        if (state>=init_area)
            return Integer.MIN_VALUE;
    }
}
```

```

int maxQaction = -1;

double max = -Integer.MIN_VALUE;
for (int i=0;i<stateDecision[state].actions.size();i++){
    // the value function of the ith action
    double valf = stateDecision[state].actions.get(i).value_function;
    if (valf>max && stateDecision[state].actions.get(i).offer.equals(sat)){
        max = valf;
        maxQaction = i;
    }
}

return maxQaction;
}

public double updatevalue(Action act,double reward){
    double next = act.value_function;
    if (nextActionE!=null)
        next = nextActionE.value_function;
    else if(nextActionG!=null)
        next =nextActionG.value_function;

    return  act.value_function + lr*(reward+disc_factor* next - act.value_function);
}

protected int maxQ( boolean next){

    int state = curr_state;
    if (next)
        state=curr_state+1;
    if (state>=init_area)
        return Integer.MIN_VALUE;

    int maxQaction = 0;

```

```

        double max = stateDecision[state].actions.get(0).value_function;
        for (int i=1;i<stateDecision[state].actions.size();i++){
            // the value function of the ith action
            double valf = stateDecision[state].actions.get(i).value_function;
            if (valf>max){
                max = valf;
                maxQaction = i;
            }
        }

        return maxQaction;
    }

```

```

@Override
int make_offer(int available_area) {
    double offer;
    if (curr_state>= init_area-1)
        return 0;

    if(available_area>satisfied)
        offer = (available_area-satisfied) * (Math.random());
    else
        offer = available_area * (Math.random()*0.5);

    double epsilon = Math.random();
    if (epsilon<0.1)
        offer = available_area * Math.random();
    return (int) Math.floor(offer);
}

```

```

@Override
Action decide_for_offer(int offer, int av_area) {

    Action maxAction;

```

```

if (nextActionG==null || nextActionL==null || nextActionE==null) {
    if (curr_state >= init_area)
        return null;
    double epsilon = Math.random();

    if (epsilon < 0.05)
        maxAction = stateDecision[curr_state].actions.get((int) Math.random() *
stateDecision[curr_state].actions.size());
    else {
        if (offer >= satisfied)
            maxAction = stateDecision[curr_state].actions.get(maxQO(false, ">=0.5"));
        else
            maxAction = stateDecision[curr_state].actions.get(maxQO(false, "<0.5"));
    }

} else{
    if (nextActionE!=null)
        maxAction = nextActionE;
    else if(offer >= satisfied)
        maxAction = nextActionG;
    else
        maxAction = nextActionL;

}

if (curr_state+2 >= init_area)
    return maxAction;
double epsilon = Math.random();

if (epsilon < 0.05) {
    nextActionE = stateDecision[curr_state + 2].actions.get((int) Math.random() *
stateDecision[curr_state].actions.size());
    nextActionG = null;
    nextActionL = null;
}

```

```

else {
    nextActionE=null;
    nextActionG = stateDecision[curr_state+2].actions.get(maxQO(false, ">=0.5"));
    nextActionL = stateDecision[curr_state+2].actions.get(maxQO(false, "<0.5"));
}
return maxAction;
}

```

```

protected void reset_state(){
    curr_state=0;
}

protected void reset_all(){
    stateDecision = new State[init_area];
    for(int i=0;i<init_area;i++){
        stateDecision[i]=new State();
        stateDecision[i].actions.add(new Action("Refuse", "<0.5",0));
        stateDecision[i].actions.add(new Action("Refuse", ">=0.5",0));
        stateDecision[i].actions.add(new Action("Accept", "<0.5",0));
        stateDecision[i].actions.add(new Action("Accept", ">=0.5",0));

    }
    curr_state=0;
}

protected void debugstate(){
    for (int i=0;i<stateDecision.length;i++){

        for (int j=0;j<stateDecision[i].actions.size();j++){
            if (i<5)
                System.out.println("State: " + i+" action: "+stateDecision[i].actions.get(j));

        }
    }
}

```

```
}
```

```
QLearningAgent.java
```

```
public class QLearningAgent extends Agent{

    public QLearningAgent(char name, int area, double satisfied, double learning_rate, double
discount_f) {
        init_area = area;
        this.satisfied=satisfied;
        lr=learning_rate;
        disc_factor=discount_f;
        stateDecision = new State[area];
        agent_name = name;
        for(int i=0;i<area;i++){
            stateDecision[i]=new State();
            stateDecision[i].actions.add(new Action("Refuse","<0.5",0));
            stateDecision[i].actions.add(new Action("Refuse",">=0.5",0));
            stateDecision[i].actions.add(new Action("Accept","<0.5",0));
            stateDecision[i].actions.add(new Action("Accept",">=0.5",0));

        }

    }

    public double updatevalue(Action act,double reward){
        double next=0;
        if (maxQ(true)>=0)
            next = stateDecision[curr_state].actions.get(maxQ(true)).value_function;

        return ((1 - lr) * act.value_function) + lr*(reward+disc_factor* next);
    }

    int maxQ(boolean next){
```

```

int state = curr_state;
if (next)
    state=curr_state+1;
if (state>=init_area)
    return Integer.MIN_VALUE;

int maxQaction = 0;

double max = stateDecision[state].actions.get(0).value_function;
for (int i=1;i<stateDecision[state].actions.size();i++){
    // the value function of the ith action
    double valf = stateDecision[state].actions.get(i).value_function;
    //System.out.println(valf + " "+ max);
    if (valf>max){
        max = valf;
        maxQaction = i;
    }
}

return maxQaction;
}

int maxQO(boolean next,String sat){

int state = curr_state;
if (next)
    state=curr_state+1;
if (state>=init_area)
    return Integer.MIN_VALUE;

int maxQaction = -1;

double max = -Integer.MIN_VALUE;
for (int i=0;i<stateDecision[state].actions.size();i++){
    // the value function of the ith action

```

```

        double valf = stateDecision[state].actions.get(i).value_function;
        if (valf>max && stateDecision[state].actions.get(i).offer.equals(sat)){
            max = valf;
            maxQaction = i;
        }
    }

    return maxQaction;
}

```

```

@Override
int make_offer(int available_area) {
    double offer;
    if (curr_state>= init_area-1)
        return 0;
    if(available_area>satisfied)
        offer = (available_area-satisfied) * (Math.random());
    else
        offer = available_area * (Math.random()*0.5);
}

```

```

double epsilon = Math.random();
if (epsilon<0.1)
    offer = available_area * Math.random();
return (int) Math.floor(offer);
}

```

```

/****

```

```

*

```

```

* @param offer

```

```

* @param av_area

```

```

* @return

```

```

*/

```

```

@Override

```

```

Action decide_for_offer(int offer, int av_area) {

    if (curr_state>=init_area)
        return null;
    double epsilon = Math.random();
    Action maxAction;
    if (epsilon<0.05)
        maxAction = stateDecision[curr_state].actions.get((int)
Math.random()*stateDecision[curr_state].actions.size());
    else {
        if (offer >= satisfied)
            maxAction = stateDecision[curr_state].actions.get(maxQO(false, ">=0.5"));
        else
            maxAction = stateDecision[curr_state].actions.get(maxQO(false, "<0.5"));
    }
    return maxAction;
}

protected void reset_all(){
    stateDecision = new State[init_area];
    for(int i=0;i<init_area;i++){
        stateDecision[i]=new State();
        stateDecision[i].actions.add(new Action("Refuse", "<0.5", 0));
        stateDecision[i].actions.add(new Action("Refuse", ">=0.5", 0));
        stateDecision[i].actions.add(new Action("Accept", "<0.5", 0));
        stateDecision[i].actions.add(new Action("Accept", ">=0.5", 0));

    }
    curr_state=0;
}

protected void reset_state(){
    curr_state=0;
}

```

```

protected void debugstate(){
    for (int i=0;i<stateDecision.length;i++){
        for (int j=0;j<stateDecision[i].actions.size();j++){
            if (i<15)
                System.out.println("State: " + i+" action: "+stateDecision[i].actions.get(j));

        }
    }
}

```

Agent.java

```

public class Environment {
    protected double reward(int offer, int newoffer,int avail_area,Agent a){
        double reward=-0.25;
        if (newoffer==-1) {
            if ( offer < a.satisfied)
                reward = -3;
            else if ( offer >= a.satisfied) {
                reward = 1;
            }
        }
        else{
            if (avail_area-newoffer < a.satisfied)
                reward = -3;
            else
                reward = 1 ;
        }

        return reward;
    }
}

```

```
public class ExternalFactors {
    double total_reward;
    int initial_area;

    public ExternalFactors(int init_area) {
        this.initial_area = init_area;
    }

    public static double calculateReward(int avail_area,int offer, int decision, Agent
agentDeciding, Agent agentOffering){
        double reward=0;
        reward+=unReward(avail_area,offer,decision, agentDeciding, agentOffering);
        reward+=usaReward(avail_area,offer,decision, agentDeciding, agentOffering);
        reward+=ukReward(avail_area,offer,decision, agentDeciding, agentOffering);
        reward+=greekReward(avail_area,offer,decision, agentDeciding, agentOffering);
        reward+=turkeyReward(avail_area,offer,decision, agentDeciding, agentOffering);

        return reward;
    }

    private static double greekReward(int avail_area,int offer, int decision, Agent agentDeciding,
Agent agentOffering){
        double r=0;
        if (agentDeciding.agent_name=='e'){
            if (decision == -1) {
                if (offer >= agentDeciding.satisfied)
                    r = 1;
                else
                    r = -5;
            }
            else {
                if (avail_area - decision >= agentDeciding.satisfied)
                    r = 1;
```

```

        else
            r = -5;
    }
}
return r;
}

```

```

private static double turkeyReward(int avail_area,int offer, int decision, Agent agentDeciding,
Agent agentOffering){
    double r=0;
    if (agentDeciding.agent_name=='t'){
        if (decision == -1) {
            if (offer >= agentDeciding.satisfied)
                r = 1;
            else
                r = -5;
        }
        else {
            if (avail_area - decision >= agentDeciding.satisfied)
                r = 1;
            else
                r = -5;
        }
    }
    return r;
}

```

/**

* decision = -1 means the agent accepted, otherwise decision value is the counteroffer

*/

```

private static double unReward(int avail_area,int offer, int decision, Agent agentDeciding,
Agent agentOffering){
    double r=0;
    // agent accepted the offer
    if (decision==-1) {
        if (offer >= agentDeciding.satisfied && avail_area - offer >= agentOffering.satisfied)
            r = 5;
    }
}

```

```

        else if (offer >= agentDeciding.satisfied) {
            r = -2;
        }
        else
            r = -1;
    }
    else {
        if (offer >= agentDeciding.satisfied && avail_area - offer >= agentOffering.satisfied) {
            r = -2;
        }
        else if (avail_area - decision >= agentDeciding.satisfied && decision >=
agentOffering.satisfied)
            // agent rejected but made a compromise offer
            r = 3;
        else //agent rejected and didn't make a compromise offer
            r = -1;
    }

    return r;
}

private static double ukReward(int avail_area,int offer, int decision, Agent agentDeciding,
Agent agentOffering){
    double r=0;
    // agent accepted the offer
    if (decision==1) {
        if (offer >= agentDeciding.satisfied && avail_area - offer >= agentOffering.satisfied)
            r = 5;
        else if (offer >= agentDeciding.satisfied)
            r = -3;
        else
            r = -2;
    }
    else {
        // agent rejected but made a compromise offer

```

```

        if (avail_area - decision >= agentDeciding.satisfied && decision >=
agentOffering.satisfied)
            r = 5;
        else //agent rejected and didn't make a compromise offer
            r = -1;
    }

    return r;
}

private static double usaReward(int avail_area,int offer, int decision, Agent agentDeciding,
Agent agentOffering){
    double r=0;
    // agent accepted the offer
    if (decision==1) {
        if (offer >= agentDeciding.satisfied && avail_area - offer >= agentOffering.satisfied)
            r = 5;
        else if (offer >= agentDeciding.satisfied)
            r = -3;
        else
            r = -2;
    }
    else {

        if (offer >= agentDeciding.satisfied && avail_area - offer >= agentOffering.satisfied)
            r=-3;
        else if (avail_area - decision >= agentDeciding.satisfied && decision >=
agentOffering.satisfied)
            // agent rejected but made a compromise offer
            r = 5;
        else //agent rejected and didn't make a compromise offer
            r = -2;
    }

    return r;
}

```

```
}
```

```
State.java
```

```
import java.util.LinkedList;
```

```
public class State {
```

```
    LinkedList<Action> actions;
```

```
    public State(){
```

```
        this.actions=new LinkedList<Action>();
```

```
    }
```

```
}
```