Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας

Διαδραστικά Συστήματα Συστάσεων υποβοηθούμενα απο τεχνικές ενισχυτικής μάθησης

Διπλωματικη Εργασια

της

ΜΑΝΔΗΛΑΡΑ ΙΩΑΝΝΑΣ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής ΕΜΠ

Αθήνα, Μάιος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας

Διαδραστικά Συστήματα Συστάσεων υποβοηθούμενα απο τεχνικές ενισχυτικής μάθησης

Διπλωματικη Εργασια

της

ΜΑΝΔΗΛΑΡΑ ΙΩΑΝΝΑΣ

Επιβλέπων: Συμεών Παπαβασιλέιου
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13 Μαιου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Συμεών Παπαβασιλέιου	Θεοδώρα Βαρβαρίγου	Γεώργιος Ματσόπουλος
Καθηγητής ΕΜΠ	Καθηγήτρια ΕΜΠ	Καθηγητής ΕΜΠ

Αθήνα, Μάιος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας

Copyright © – All rights reserved. Με την επιφύλαξη παντός δικαιώματος.
Μανδηλαρά Ιωάννα, 2022.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα.

Το περιεχόμενο αυτής της εργασίας δεν απηχεί απαραίτητα τις απόψεις του Τμήματος, του Επιβλέποντα, ή της επιτροπής που την ενέκρινε.

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Μανδηλαρά Ιωάννα

13 Μαιου 2022

Περίληψη

Στις μέρες μας, η ανάπτυξη των Κοινωνικών και Συναισθηματικών ικανοτήτων των μαθητών είναι ζωτικής σημασίας λόγω του σημαντικού μέρους του χρόνου που περνούν στο σχολείο κατά την διάρκεια ευαίσθητων περιόδων ανάπτυξης της συναισθηματικής νοημοσύνης τους. Οι κοινωνικο-συναισθηματικές δεξιότητες βοηθούν τους μαθητές να διαχειρίζονται με επιτυχία την καθημερινή τους ζωή βελτιώνοντας τις μαθησιακές και κοινωνικές δυσκολίες τους . Αυτό προτρέπει στην δημιουργία εργαλείων που θα βοηθούν τους καθηγητές να ολοκληρώνουν δραστηριότητες κοινωνικής και συναισθηματικής μάθησης και να αξιολογούν τον αντίκτυπο που επιτυγχάνεται. Ως εκ τούτου, για την αντιμετώπιση των προαναφερθείσων αναγκών, ο συνδυασμός συστημάτων συστάσεων με τεχνικές μηχανικής μάθησης μπορεί να αποδειχθεί ευεργετικός στη δημιουργία έξυπνων και αυτο-εκπαιδευτικών εργαλείων, ικανών να προτείνουν δραστηριότητες που εστιάζουν στις κοινωνικές και συναισθηματικές ανάγκες των εκπαιδευτικών ομάδων. Ειδικότερα, τα συστήματα συστάσεων που βασίζονται στην Ενισχυτική Μάθηση (RL) έχουν γίνει ένα αναδυόμενο ερευνητικό θέμα τα τελευταία χρόνια. Το γεγονός ότι η βαθιά ενισχυτική μάθηση αξιοποιεί τη ικανότητα μάθησης των βαθιών νευρωνικών δικτύων για την αντιμετώπιση προβλημάτων, που ήταν πολύ περίπλοκα για τις κλασικές τεχνικές RL, οδηγεί στην ανάπτυξη βελτιωμένων διαδραστικών συστημάτων συστάσεων. Στο τρέχον χειρόγραφο, περιγράφουμε λεπτομερώς ένα σύστημα συστάσεων βασισμένο σε RL που στοχεύει στην σύσταση εκπαιδευτικών δραστηριοτήτων στους δασκάλους, προκειμένου να βελτιώσει τις κοινωνικο-συναισθηματικές ικανότητες των μαθητών, αξιοποιώντας τους αλγόριθμους Deep RL. Στη συνέχεια, παρουσιάζουμε τα αποτελέσματα εκπαίδευσης και αξιολόγησης αυτού του διαδραστικού συστήματος συστάσεων, που παράγονται από την εφαρμογή αλγορίθμων Deep RL, όπως Advantage Actor to Critic, Trust Region Policy Optimization και Proximal Policy Optimization.

Λέξεις Κλειδιά

Σύστημα σύστασης, Μηχανική Μάθηση, Βαθιά Ενισχυτική Μάθηση, Αλγόριθμοι ενισχυτικής μάθησης, Policy Gradient μέθοδοι, Εκπαίδευση, Κοινωνικές και Συναισθηματικές ικανότητες

Abstract

Nowadays, the development of Social and Emotional competences of students is vital due to the significant portion of time they spend in school during sensitive periods of brain development. Socio-emotional skills help students successfully manage everyday life by improving their learning and social outcomes. This prompts for tools to help tutors in accomplishing social and emotional learning activities and evaluating the impact achieved. Therefore, to address the aforementioned needs, the combination of recommendation systems with machine learning techniques can prove beneficial in creating intelligent and self-learning tools capable of recommending activities focusing on the social and emotional needs of educational groups. In particular, Reinforcement Learning (RL) based recommender systems have become an emerging research topic in recent years. The fact that deep reinforcement learning leverages the learning capacity of deep neural networks to tackle problems that were too complex for classic RL techniques lead to the development of enhanced interactive recommender systems. In the current manuscript, we detail an RL based recommender system that aims to recommend educational activities to teachers, in order to improve the social-emotional competences of students, taking advantage of Deep RL algorithms. Then, we present the results of training and evaluation of this interactive recommendation system, produced by the implementation of Deep RL algorithms, such as Advantage Actor to Critic, Trust Region Policy Optimization and Proximal Policy Optimization.

Keywords

Recommendation systems, Machine Learning, Deep Reinforcement Learning, Reinforcement Learning (RL) based recommender systems,Reinforcement Learning algorithms, Policy Gradient methods, Education, Social and Emotional competences

στην οικογένεια μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Συμεών Παπαβασιλείου για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Διαχείρισης Δικτύων και Βέλτιστου Σχεδιασμού. Επίσης ευχαριστώ ιδιαίτερα τον Αναστάσιο Ζαφειρόπουλο και την Ελένη Φωτοπούλου, ερευνητές στο εργαστήριο, για την καθοδήγησή τους, την συνεχή υποστήριξη τους και την εξαιρετική συνεργασία που είχαμε. Τέλος, θα ήθελα να ευχαριστήσω θερμά την οικογένεια μου, τον σύντροφο μου και τους φίλους μου για την καθοδήγηση, την στήριξη τους και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Μάιος 2022

Μανδηλαρά Ιωάννα

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πρόλογος

Η παρούσα διπλωματική εργασία διενεργήθηκε κατα το χειμερινό εξάμηνο του Ακαδημαικού Έτους 2021-2022, στα πλαίσια του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών στο επιστημονικό πεδίο ¨Επιστήμη Δεδομένων και Μηχανική Μάθηση¨ του Εθνικού Μετσόβιου Πολυτεχνείου.

Η εργασία πραγματοποιήθηκε στο εργαστήριο Διαχείρισης Δικτύων και Βέλτιστου Σχεδιασμού, του Εθνικού Μετσόβιου Πολυτεχνείου.

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της διπλωματικής

Στις μέρες μας, η ανάπτυξη τόσο των κοινωνικών όσο και των συναισθηματικών ικανοτήτων των μαθητών είναι μια απο τις πιο ενδιαφέρουσες προκλήσεις που καλούνται να αντιμετωπίσουν οι καθηγητές μιας εκπαιδευτικής ομάδάς. Η βελτίωση αυτών των ικανοτήτων τόσο των μαθητών όσο και των καθηγητών μπορεί να επιφέρει αποτελέσματα σωστής μάθησης και κοινωνικής αλληλεπίδρασης, όπως επίλυση προβλημάτων, δημιουργία καλού και θετικού κλίματος αλληλεπιδράσεων στην τάξη, διαχείριση άγχους και πρόληψη σοβαρών προβλημάτων συμπεριφοράς σε νεότερες ηλικίες. Έχουν γίνει προσπάθειες με το πέρασμα των χρόνων να συμπεριληφθούν εκπαιδευτικές δραστηριότητες και παρόλ' αυτά μέχρι και σήμερα χρειάζονται περαιτέρω βελτίωση αυτές οι ικανότητες. Επιπλέον η υιοθέτηση και διεξαγωγή τέτοιου είδους δραστηριοτήτων δεν είναι τόσο απλή, λόγω της απουσίας αξιολογήμενου υλικού και της ανεπαρκούς τεχνογνωσίας(σε ορισμένες περιπτώσεις) απο την πλευρά των εκπαιδευτικών. Ειδικότερα τα τελευταία 2 χρόνια, με την έξαρση της πανδημίας, δεν μπορούσαν να αναπτυχθούν δραστηριότητες με φυσική παρουσία, με αποτελέσμα να είναι υπαρκτή η ανάγκη αντιμετώπισης τέτοιου είδους προβλημάτων, που απαιτούν μια διαδικτυακή επικοινωνία μεταξύ μαθητών και δασκάλων.

Επομένως, ύπαρχει ανάγκη ανάπτυξης εργαλείων που στοχεύουν στη διευκόλυνση των δασκάλων να πραγματοποιήσουν εκπαιδευτικές δραστηριότητες που μπορούν να βελτιώσουν τα κοινωνικά και συναισθηματικά χαρακτηριστικά των μαθητών τους, καθώς την αξιολόγηση τους σε ατομικό επίπεδο. Ειδικά στην τωρινή εποχή, με την ανάπτυξη της τεχνολογίας, η μηχανική μάθηση μπορεί να υποβοήθησει και αυτόν τον τομέα και να συνδράμει σημαντικά σε καθημερίνα προβλήματα της εκπαιδευτικής διαδικασίας. Προς αυτή την κατεύθυνση, η εκμετάλλευση των τεχνολογικών λύσεων που προτείνονται απο τα συστήματα σύστασης και τη μηχανική μάθηση φαινεται τρομέρα υποσχόμενη. Ο συνδυασμός αυτών των δύο τεχνολογίων μπορεί να προσφέρει στοχευμένες συστάσεις για τη βελτίωση των κοινωνικών και των συναισθηματικών ικανοτήτων των μαθητών και να εντοπίσει τις ανάγκες μιας εκπαιδευτικής ομάδας. Η ανάπτυξη ενός συστήματος σύσταση, υποβοηθούμενο απο τεχνικές μηχανικής μάθησης, μπορεί να αποτελέσει ένα εργαλείο ουσιώδους σημασίας για τους καθηγητές.

Πιο συγκεκριμένα, το είδος της μηχανικής μάθησης που μπορεί να συνδυαστεί με τα συστήματα σύστασης, είναι η Ενισχυτική Μάθηση. Ειδικά, τα τελευταία χρόνια, η ενισχυτική μάθηση χρησιμοποιείται σε διάφορους τομείς και έχει γίνει ευρέως γνωστή. Αξιοποιώντας

λοιπόν τις τεχνικές της ενισχυτικής μάθησης, η δυναμική της αλληλεπίδρασης των μαθητών μεταξύ τους μπορεί να μοντελοποιηθει, να παρακολουθηθεί και να αξιολογηθεί και να προταθούν οι ανάλογες δραστηριότητες που στοχεύουν στη βελτίωση και εξέλιξη της εκπαιδευτικής ομάδας, βάσει των κοινωνικών και συναισθηματικών ικανοτήτων της. Στη συγκεκριμένη διπλωματική εργασία, περιγράφεται λεπτομερώς ένα περιβάλλον για ένα σύστημα σύστασης με τη βοήθεια της ενισχυτικής μάθησης, το οποίο είναι ικανό να παρέχει συστάσεις στους καθηγητές.

Το πρωταρχικό λοιπόν βήμα για την ανάπτυξη συστημάτων σύστασης που αποσκοπούν στην ενίσχυση των κοινωνικών και εκπαιδευτικών ικανοτήτων μιας εκπαιδευτικής ομάδας αφορά την ομοιογένεια που πρέπει να υπάρχει στην αναπαράσταση των πληροφοριών που συλλέγονται με βάση την υιοθέτηση και την προσαρμογή μοντέλων Συναισθηματικής Νοημοσύνης. Με τον όρο της Συναισθηματικής Νοημοσύνης αναφερόμαστε στην ικανότητα των άτομων να μπορούν να αναγνωρίσουν τα δικά τους συναισθήματα αλλά και τα συναισθήματα των άλλων ανθρώπων, να διακρίνουν τα διαφορετικά συναισθήματα και να τα κατηγοροποιούν κατάλληλα, να χρησιμοποιούν πληροφορίες βάσει της συναισθηματικής τους νοημοσύνης για να καθοδηγούν τη σκέψη τους και τη συμπεριφορά τους και να διαχειρίζονται τα συναισθήματα τους για να μπορούν να προσαρμοστούν στο περιβάλλον τους ή να επιτύχουν τους στόχους τους [13].

Τα τρία κύρια μοντέλα προς εξέταση είναι το μοντέλο της ικανότητας [14], το μικτό μοντέλο [15] και το μοντέλο των χαρακτηριστικών [16]. Σε όλες τις περιπτώσεις, ο ορισμός της Συναισθηματικής Νοημοσύνης και η διασύνδεση των δεικτών της με συγκεκριμένες συναισθηματικές ικανότητες (είτε διδάσκονται είτε βελτιώνονται με εκπαιδευτικές δραστηριότητες) δεν ειναι αυστήρος, συγκεκριμένος ή τυποποιημένος.

Το μοντέλο βασισμένο στο [10] εκφράζει τις συναισθηματικές ικανότητες ως ενα σύνολο απο ικανότητες μικρής κλίμακας που μπορούν να αξιολογηθούν και παρέχει κατέπέκταση μια πολύ ενδιαφέρουσα οπτική για τη μοντελοποίηση του προβλήματος.

1.2 Οργάνωση του τόμου

Η εργασία αυτή είναι οργανωμένη σε 6 κεφάλαια: Στο Κεφάλαιο 2 δίνεται το θεωρητικό υπόβαθρο των βασικών τεχνολογιών που σχετίζονται με τη διπλωματική αυτή. Αρχικά περιγράφονται τα συστήματα σύστασης στην εκπαίδευση και με τη χρήση της μηχανικής μάθησης και πιο συγκεκριμένα της ενισχυτικής μάθησης και στη συνέχεια περίγραφονται οι αλγόριθμοι που χρησιμοποιούνται στο πλαίσιο της διπλωματικής εργασίας.Στο Κεφάλαιο 3 περιγράφεται η σχεδίαση του διαδραστικού συστήματος σύστασης και έπειτα στο Κεφάλαιο 4 παρουσιάζονται λεπτομερώς τα εργαλεία που χρησιμοποιούνται και πιο τεχνικές λεπτομέρειες υλοποίησης του συστήματος σύστασης βάσει αυτών των εργαλείων. Στο Κεφάλαιο 5 παρουσιάζεται η μεθοδολογία που ακολουθείται και τα αποτελέσματα που προκύπτουν. Τέλος, κλείνοντας, στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα που προκύπτουν και οι μελλοντικές επεκτάσεις της διπλωματικής εργασίας.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

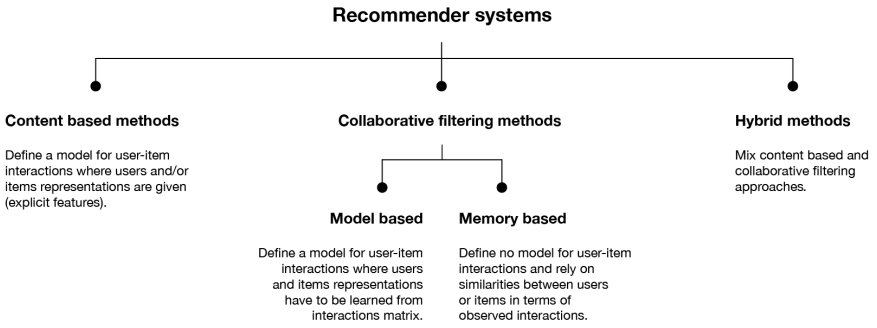
Στο κεφάλαιο αυτό παρουσιάζονται αναλυτικά η θεωρητική προσέγγιση της συγκεκριμένης εργασίας, δηλαδή την διαφορά ανάμεσα στα διαδραστικά και μη συστήματα σύστασης. Με τον όρο διαδραστικά, όπως προαναφέρεται, εννοούμε με την βοήθεια της μηχανικής μάθησης και το είδος των αλγορίθμων που χρησιμοποιήθηκαν.

2.1 Συστήματα Σύστασης

Τις τελευταίες δεκαετίες, με την συνεχή άνοδο υπηρεσιών όπως το Youtube, το Netflix και παρόμοιες διαδικτυακές υπηρεσίες, τα συστήματα συστάσεων έχουν κυριαρχήσει όλοένα και περισσότερο στη ζωή μας. Απο τη διαδικτυακή διαφήμιση (σύσταση του κατάλληλου περιεχομένου στους χρήστες, βάσει των προτιμήσεων τους) μέχρι το διαδικτυακό εμπόριο (σύσταση άρθρων σε αγοραστές, βάσει των ενδιαφερόντων τους), τα συστήματα συστάσεων ειναι πλέον κυρίαρχα στην καθημερινή διαδικτυακή ζωή μας.

2.1.1 Μεθοδοι συστημάτων σύστασης

Η παραδοσιακή αυτή εκδοχή των συστημάτων συστάσεων κατηγοροποιείται σε 3 μεθόδους: content-based, collaborative filtering and hybrid systems[17] [18]. Οι Collaborative filtering μέθοδοι είναι εκείνες οι οποίες βασίζονται σε παλιότερες αλληλεπιδράσεις μεταξύ χρηστών και αντικειμένων, αναγνωρίζουν ομοιότητες μεταξύ των χρηστών και συστήνουν νέα αντικείμενα βάσει των ομοιοτήτων απο παρόμοιους χρήστες. Σε αντίθεση, οι Content-based μέθοδοι εισάγουν επιπρόσθετη πληροφορία για τους χρήστες και τα αντικείμενα και συνεπώς οι συστάσεις πραγματοποιούνται βάσει των χαρακτηριστικών που προσδιορίζουν τους χρήστες χρησιμοποιώντας παράλληλα πληροφορία απο το περιεχόμενο των παλιότερων αξιολογημένων αντικειμένων. Ο συνδυασμός των δύο παραπάνω μεθόδων αποτελεί την 3η μέθοδο, δηλαδή το hybrid system, η οποία επιλέγει τους καλύτερους αλγόριθμους με σκοπό την επίτευξη καλύτερης αποδοτικότητας και την αποφυγή περιορισμένων που εισάγει ο κάθε αλγόριθμος απο μόνος του.



Σχήμα 2.1: *Κατηγορίες μεθόδων συστημάτων σύστασης [1]*

2.1.2 Συστήματα σύστασης στην εκπαίδευση

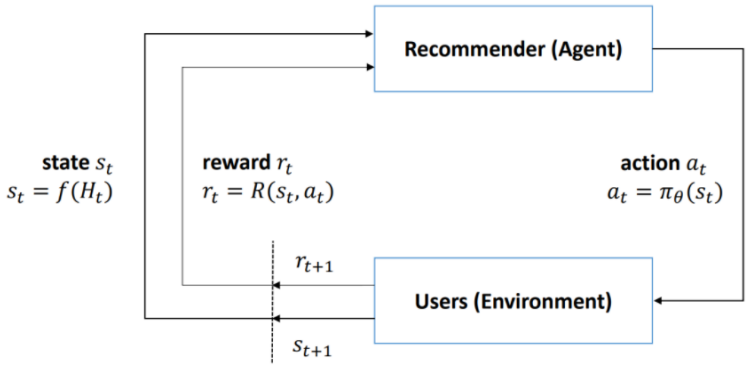
Παρόλο που τα συστήματα σύστασης χρησιμοποιούνται ευρέως στο τομέα του e-commerce, εγκειται η ανάγκη εφαρμογής τους σε τελείως διαφορετικούς τομείς, όπως η εκπαίδευση. Ο ρόλος τους στην εκπαίδευση ειναι ουσιώδης, μιας και μπορεί να θεωρηθεί ως ενα εργαλείο του διδάσκοντα για να βελτιωθεί η εκπαιδευτική διαδικασία και προταθούν κατάλληλες δραστηριότητες διδασκαλίας και μάθησης με στόχο την καλύτερη σχέση μαθητή με καθηγητή. Οι προαναφερθείσες μέθοδοι έχουν εφαρμοστεί στον εκπαιδευτικό τομέα, με πιο συχνή την υβριδική, και στοχεύουν στη σύσταση μαθησιακού υλικού βάσει τον ενδιαφερόντων και της μαθησιακής διαδικτυακής διαδρομής του μαθητή, στη σύσταση δραστηριοτήτων διδασκαλίας και μάθησης και συνολικά σε μια βελτιστοποιημένη ακαδημαική απόδοση [19] [20].

2.1.3 Συστήματα σύστασης με μηχανική μάθηση

Παρόλο που η παραδοσιακή εκδοχή των συστημάτων σύστασης έχει χρησιμοποιηθεί σε πολλούς διαφορετικούς τομείς συμπεριλαμβανομένων της σύστασης ταινιών, μουσικής, περιεχομένου και διαδικτυακών μαθησιακών μονοπατιών, με την τεράστια άνοδο της μηχανικής μάθησης εχει εμφανιστεί μια νέα τάση σχετικά με τα συστήματα σύστασης, τα λεγόμενα διαδραστικά συστήματα σύστασης. Η βαθιά μηχανική μάθηση, δηλαδή πχ. η χρήση βαθιών νευρωνικών δικτύων, όπως τα συνελειτικτικά και τα αναδρομικά νευρωνικά δίκτυα, έπαιξε σημαντικό ρόλο στην εξέλιξη της κλασσικής εκδοχής των συστημάτων σύστασης, δίνοντας δυνατότητες εύρεσης πιο πολύπλοκων και μη γραμμικών σχέσεων μεταξύ των χρηστών και των αντικειμένων και παράλληλα αύξησης της αποδοτικότητας. Όμως, γνωρίζουμε πολύ καλά ότι τέτοιου είδους μοντέλα είναι γνωστά ως black-boxes και συνεπώς είναι δύσκολα ερμηνεύσιμα, παράλληλα απαιτούν μεγάλο όγκο δεδομένων για την παραμετροποίηση τους και κοστίζουν υπολογιστικά [21]. Παράλληλα, η κλασσική εκδοχή των συστημάτων σύστασης προσφέρει μια στατική λίστα απο συστάσεις στους χρήστες και το feedback που λαμβάνουν περιορίζεται σε αποδοχή ή όχι της σύστασης που προτάθηκε, δηλαδή επικεντρώνεται στην άμεση απάντηση του χρήστη χωρίς να λαμβάνονται υπόψιν οι μακροπρόθεσμενες επιδράσεις στην επακόλουθη συμπεριφορά του[22]. Καπως ετσι εμφανίζονται τα διαδραστικά συστήματα σύστασης, δηλαδή εκείνα με τη βοήθεια ενισχυτικής μάθησης, τα οποία μπορούν να χειριστούν τέτοιου είδους επιδράσεις.

2.1.4 Διαδραστικά συστήματα σύστασης

Η ενισχυτική μηχανική μάθηση είναι ένα πεδίο της μηχανικής μάθησης που μελετά προβλήματα, στα οποία ο πράκτορας μαθαίνει τι πρέπει να κάνει για να μεγιστοποιήσει μια ανταμοιβή (reward) μέσω της αλληλεπίδρασης του με το περιβάλλον. Η ικανότητα ενός πράκτορα ενισχυτικής μάθησης να μαθάινει μέσω της ανταμοιβής του απο το περιβάλλον χωρίς παράλληλα την απαίτηση δεδομένων εκπαίδευσης ταιριάζει απόλυτα σε ένα σύστημα σύστασης. Δεδομένου ότι το πρόβλημα της σύστασης των κατάλληλων αντικειμένων στους χρήστες δεν αποτελεί μόνο ένα πρόβλημα πρόβλεψης(παραδοσιακή εκδοχή συστημάτων σύστασης) αλλα ένα πρόβλημα διαδοχικής απόφασης, τότε μια μαρκοβιανή αλυσίδα απόφασης αποτελεί το καταλληλότερο μοντέλο για το συστήματα σύστασης[23]. Συνεπώς συγκρίνοντας την κλασσική εκδοχή των συστημάτων σύστασης με μια μαρκοβιανή αλυσίδα που επιλύεται με μεθόδους ενισχυτικής μάθησης, τότε ο αλγόριθμος του κλασσικού συστήματος σύστασης είναι ανάλογος με τον πράκτορα ενισχυτικής μάθησης, η μεγιστοποίηση της ικανοποίησης του χρήση είναι ανάλογη με μεγιστοποίηση της ανταμοιβής του πράκτορα και οτιδήποτε εκτός του πράκτορα,δηλαδή οι χρήστες και τα αντικείμενα, μπορούν να θεωρηθούν σαν το περιβάλλον.



Σχήμα 2.2: *Διαδραστικά συστήματα σύστασης [2]*

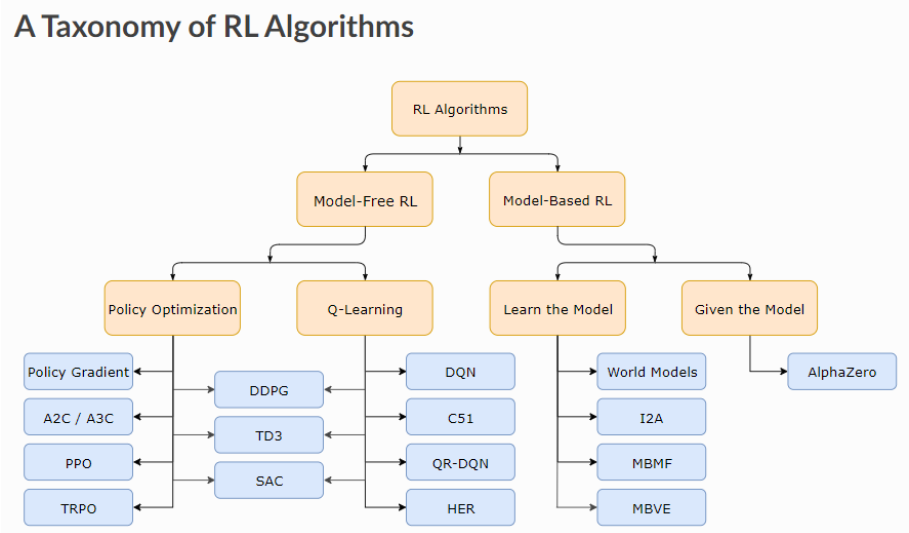
Βάσει της μαρκοβιανής ιδιότητας,τα προβλήματα σύστασης με ενισχυτική μάθηση θα προτείνουν μια νέα σύσταση ανεξάρτητα απο τις προηγουμένες συστάσεις. Πιο συγκεκριμένα, αυτό δίνει το πλεονέκτημα της ισορροπίας ανάμεσα σε exploration και exploitation, δηλαδή ο πράκτορας δεν προτείνει στους χρήστες μόνο το πιο χρήσιμο περιεχόμενο αλλα προτείνει και τυχαίο περιεχομένο, προκαλώντας νέο πιθανό ενδιαφέρον σε αυτούς. Ένα απο τα σπουδαιότερα πλεονεκτήματα που δίνουν τα μοντέλα ενισχυτικής μάθησης είναι ότι συνεχώς μαθαίνουν, δηλαδή ότι όσο αλλάζουν τα ενδιαφέροντα του χρήστη τόσο αλλάζει και το προτεινόμενο περιεχόμενο, με αποτέλεσμα να προκύπτει ένα πιο εύρωστο μοντέλο.[24] Συνδυάζοντας όλα τα παραπάνω, ένα διαδραστικό σύστημα σύστασης δίνει στον χρήστη τη δυνατότητα να λάβει μέρος στη διαδικασία επιτρέποντας του να αλληλεπίδρασει με τις συστάσεις που προτείνονται , να δώσει το απαραίτητο feedback και να επηρεάσει τα αποτελέσματα σε πραγματικό χρόνο[25].

2.2 Είδη αλγορίθμων ενισχυτικής μάθησης

Ο στόχος της ενισχυτικής μάθησης είναι να μάθει ο πράκτορας μια καλή στρατηγική μέσω πειραματικών δοκιμών και της σχετικού feedback που λαμβάνει. Βάσει αυτής της βέλτιστης στρατηγικής, ο πράκτορας είναι ικανός να προσαρμοστεί στο περιβάλλον και να μεγιστοποιήσει τις μελλοντικές του ανταμοιβές. Για την προσέγγιση του βασικού αυτού στόχου, υπάρχουν πολλά διαφορετικά είδη αλγορίθμων 2.3. Το βασικό ερώτημα που προκύπτει είναι αν ο πράκτορας έχει πρόσβαση, ή αλλιώς μαθαίνει το μοντέλο του περιβάλλοντος. Με τον όρο μοντέλο, εννοούμε μια συνάρτηση, η οποία προβλέπει τις μεταβάσεις ανάμεσα στις καταστάσεις(state-transitions) και τις ανταμοιβές [26]. Οπότε προκύπτουν τα δύο βασικά είδη αλγορίθμων:

- **Model-Based RL**: Οι συγκεκριμένοι αλγόριθμοι βασίζονται στο μοντέλο του περιβάλλοντος, δηλαδή ο πράκτορας είτε μαθαίνει ρητά το μοντέλο είτε μαθαίνει μέσω του μοντέλου. Πιο συγκεκριμένα, η Model-based μάθηση προσπαθεί να μοντελοποιήσει το περιβάλλον και στη συνέχεια επιλέγει τη βέλτιστη στρατηγική βάσει του μοντέλου που έχει μάθει [27].

- **Model-Free RL**: Σε αυτή την περίπτωση δεν υπάρχει εξάρτηση απο το μοντέλο κατά τη μάθηση, δηλαδή ο πράκτορας μαθαίνει τη στρατηγική που μπορεί να επιτύχει τη βέλτιστη συμπεριφορα, χωρίς όμως τη χρήση του μοντέλου, δηλαδή μέσω δοκιμών και σφαλμάτων [27].



Σχήμα 2.3: *Ταξινόμηση των αλγορίθμων ενισχυτικής μάθησης [3]*

Γενικότερα, οι Model-Free προσεγγίσεις επιλέγονται περισσότερο έναντι των προσεγγίσεων βασισμένων στο μοντέλο του περιβάλλοντος λόγω της πιο εύκολης υλοποίησης τους, της ταχύτερης σύγκλισης και της απλότητας τους. Έχουν αναπτυχθεί και δοκιμαστεί περισσότερο και είναι πιο εύκολο να υλοποιηθούν αλλα και να βελτιστοποιηθούν οι παράμετροι τους [26] [28]. Είναι οι πιο κατάλληλες για τη συγκεκριμένη εργασία λόγω του ότι οι συγκεκριμένοι αλγόριθμοι μπορούν να εφαρμοστούν για διαφορετικούς τύπους του χώρου ενεργειών (action

space) και συνεπώς αυτή η ιδιαίτεροτητα τους ταιριάζει απόλυτα στο προβλήμα σύστασης με ενισχυτική μάθηση που θέλουμε να υλοποιήσουμε.

2.2.1 Policy Optimization Αλγόριθμοι

Πιο συγκεκριμένα, για το πρόβλημα μας χρησιμοποιούνται οι μέθοδοι βελτιστοποίησης στρατηγικής, λόγω του ότι οι αλγόριθμοι που τις απαρτίζουν μπορούν να δουλέψουν με πολυδιάστατους δυαδικούς χώρους ενεργειών, τους οποιούς χρησιμοποιούμε για την υλοποίηση του συστήματος σύστασης.

Σε αυτή την οικογένεια μεθόδων, ο πράκτορας μαθαίνει απευθείας τη στρατηγική με μια παραμετροποιημένη συνάρτηση ως προς ϑ, $\pi_\vartheta(a|s)$. Στόχος τους ειναι η βελτιστοποίηση της παραμέτρου ϑ είτε απευθείας μέσω του gradient ascent της συνάρτησης αναμενόμενης ανταμοιβής (expected reward) $J(\pi_\vartheta)$ είτε έμμεσα μεγιστοποιώντας τις τοπικές προσεγγίσεις της συνάρτησης ανταμοιβής. Μέσω του gradient ascent, η παράμετρος ϑ ωθείται ως προς την κατεύθυνση που συστήνεται απο το $\nabla J(\pi_\vartheta)$ και με αυτόν τον τρόπο μπορεί να βρεθεί η βέλτιστη τιμή της παραμέτρου ϑ για την πολιτική $\pi_\vartheta(a|s)$, η οποία οδηγεί στην υψηλότερη ανταμοιβή (return) [26] [7].

Σύμφωνα με το Policy Gradient θεωρήμα, προκύπτει το παρακάτω αποτέλεσμα, το οποίο αποτελεί τη θεωρητική βάση για όλους τους αλγορίθμους αυτής της κατηγορίας[29].

$$\nabla J(\pi_\vartheta) = \nabla E_{\pi_\vartheta}[r(\tau)] = E_{\pi_\vartheta}[r(\tau)\nabla log(\pi_\vartheta)] \tag{2.1}$$

όπου $r(\tau)$ είναι η συνολική ανταμοιβή.

Η παραπάνω έκφραση μπορεί να πάρει διαφορετικές μορφές για κάθε αλγόριθμο και την υλοποίηση του 2.4, σύμφωνα με [4] .

Policy gradient methods maximize the expected total reward by repeatedly estimating the gradient $g := \nabla_\theta \mathbb{E}\left[\sum_{t=0}^{\infty} r_t\right]$. There are several different related expressions for the policy gradient, which have the form

$$g = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right], \tag{1}$$

where Ψ_t may be one of the following:

1. $\sum_{t=0}^{\infty} r_t$: total reward of the trajectory.
2. $\sum_{t'=t}^{\infty} r_{t'}$: reward following action a_t.
3. $\sum_{t'=t}^{\infty} r_{t'} - b(s_t)$: baselined version of previous formula.
4. $Q^\pi(s_t, a_t)$: state-action value function.
5. $A^\pi(s_t, a_t)$: advantage function.
6. $r_t + V^\pi(s_{t+1}) - V^\pi(s_t)$: TD residual.

The latter formulas use the definitions

$$V^\pi(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \qquad Q^\pi(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \tag{2}$$

$$A^\pi(s_t, a_t) := Q^\pi(s_t, a_t) - V^\pi(s_t), \quad \text{(Advantage function)}. \tag{3}$$

Σχήμα 2.4: *Γενική μορφή των policy gradient μεθόδων [4]*

Οι αλγόριθμοι που χρησιμοποιούνται στα πλαίσια αυτής της εργασίας είναι ο Advantage Actor to Critic(A2C), Proximal Policy Optimization(PPO) και Trust Region Policy Opti-

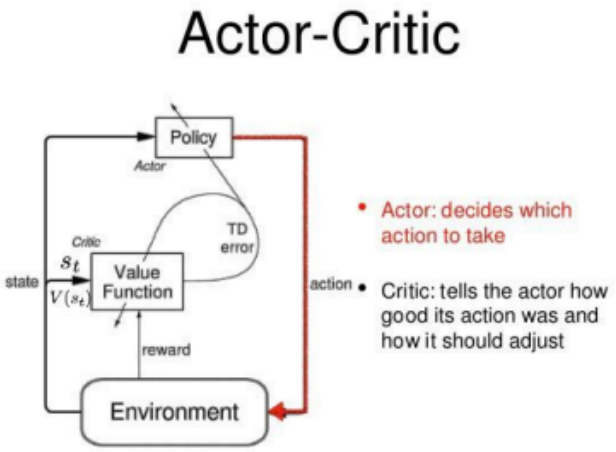
mization (TRPO). Οι 3 συγκεκριμένοι αλγόριθμοι βασίζονται στην ιδέα του Actor-Critic και διάφερουν ως προς την ανανέωση της παραμέτρων της πολιτικής του actor. Τα δύο βασικά στοιχεία απο τα οποία αποτελείται είναι το μοντέλο στρατηγικής (policy model) και η value συνάρτηση, η οποία δηλώνει πόσο καλή είναι η κατασταση στην οποία βρισκόμαστε. Συνεπώς οι Actor-Critic μέθοδοι αποτελούνται απο 2 μοντέλα:

- **Actor**: Δέχεται σαν είσοδο την κατάσταση και δίνει σαν έξοδο την καλύτερη ενέργεια, δηλαδή ανανεώνει την παράμετρο θ της $\pi_\theta(a|s)$, στην κατεύθυνση που προτείνει ο Critic. Ελέγχει δηλαδή πως συμπεριφέρεται ο πράκτορας μαθαίνοντας την καλύτερη στρατηγική. Μπορεί να είναι ενα πλήρως συνδεδομένο νευρωνικό ή ένα συνελικτικό δίκτυο.

- **Critic**: Δέχεται σαν είσοδο το περιβάλλον και την ενεργεια απο τον actor και επιστρέφει value συνάρτηση. Αξιολογεί δηλαδή την ενέργεια υπολογίζοντας τη value συνάρτηση. Μπορεί να είναι ενα πλήρως συνδεδομένο νευρωνικό ή ένα συνελικτικό δίκτυο.

Αυτά τα δύο μοντέλα αλληλεπιδρούν μεταξύ τους και προσπαθούν να γίνουν καλύτερα, βελτιώνοντας τον ρόλο τους. Η εκπαίδευση των δύο δικτύων γίνεται ξεχωριστά και χρησιμοποιείται gradient ascent για την ενημέρωση των βάρων τους. Όσο περνάει ο καιρός, ο actor μαθαίνει να παράγει όλο και καλύτερες δράσεις (αρχίζει να μαθαίνει τη στρατηγική) και ο critic γίνεται όλο και καλύτερος στην αξιολόγηση αυτών των ενεργειών [7] [30].

Ένα εξαιρετικό παράδειγμα, σύμφωνα με [30], είναι το εξής: Ενα παιδί (Actor) δοκιμάζει διαρκώς νέα πράγματα και εξερευνεί το περιβάλλον γύρω του, όπως πχ. δαγκώνει τα παιχνίδια του. Η μητέρα του (Critic) τον παρακολουθεί και είτε τον επικρίνει είτε τον ενθαρρύνει. Το παιδί ακούει τη μητέρα του και προσαρμόζει τη συμπεριφορά του και όσο μεγαλώνει μαθαίνει ποιες ενέργειες είναι καλές ή κακές και ουσιαστικά μαθαίνει το παιχνίδι που ονομάζεται ζωή.

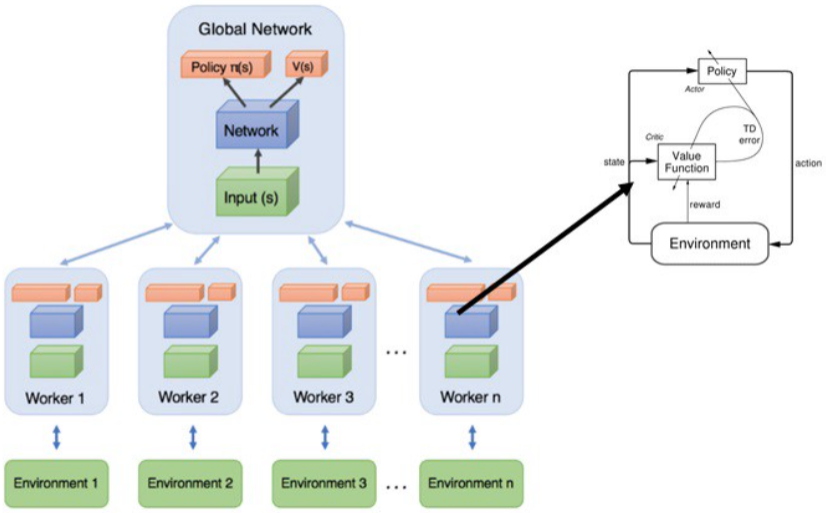


Σχήμα 2.5: *Actor-Critic [5]*

2.2.2 Advantage Actor to Critic (A2C)

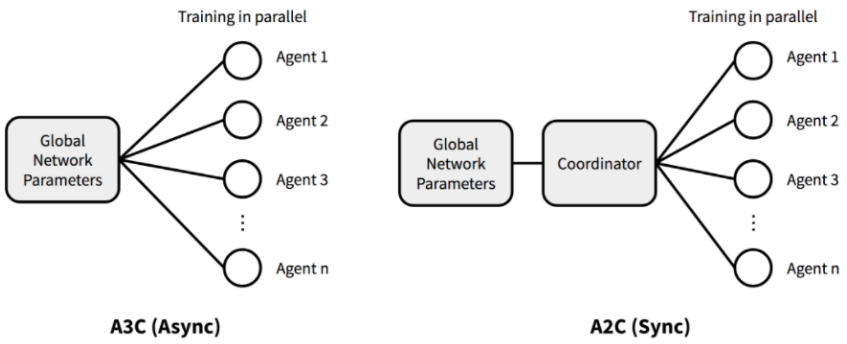
Ο Advantage Actor to Critic είναι μια συγχρονισμένη και ντετερμινιστική εκδοχή του A3C. Αρχικά ονομάζεται Advantage διότι σύμφωνα με την 2.4 χρησιμοποιεί την Advantage

συνάρτηση, η οποία απαντάει στην ερώτηση: πόσο καλύτερη είναι μια ενέργεια σε σύγκριση με όλες τις άλλες σε μια συγκεκριμένη κατάσταση ή πόσο καλύτερη θα ήταν η ανταμοιβή για ένα πράκτορα απο ο,τι θα περιμένε; [30]. Σε αυτή την περίπτωση, ο Critic επιστρέφει τις τιμές της Advantage συνάρτησης, δηλαδή η αξιόλογηση της ενέργειας βασίζεται πλεον στο πόσο καλύτερη μπορεί να γίνει και όχι στο πόσο καλή είναι. Όσον αφορά τον A3C, ο A3C χρησιμοποιεί πολλούς ανεξάρτητους πράκτορες (νευρωνικά δίκτυα με τα δικά τους βάρη) που αλληλεπιδρούν με ένα διαφορετικό αντίγραφο του περιβάλλοντος παράλληλα. Αυτοί οι πράκτορες εκπαιδεύονται παράλληλα και ενημερώνουν **ασυγχρόνιστα** ένα κεντρικό δίκτυο, που έχει τις κοινές παραμέτρους και μετά απο κάθε δικιά τους ενημέρωση ενημερώνουν τις δικές τους παραμέτρους στις κοινές [31].



Σχήμα 2.6: *A3C [6]*

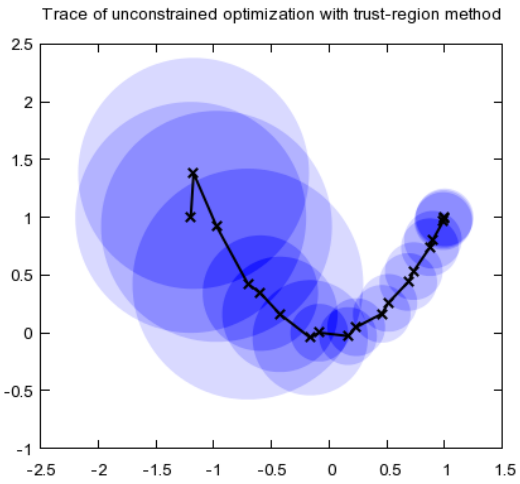
Ο A2C λοιπόν αποτελεί μια συχρονισμένη εκδοχή του A3C [31],διότι χρησιμοποιεί πολλούς ανεξάρτητους πράκτορες με τα δικά τους βάρη, οι οποίοι αλληλεπιδρούν με ένα αντίγραφο του περιβάλλοντος, εκπαιδεύονται παράλληλα, στη συνέχεια ενημερώνουν **παράλληλα** ένα κεντρικό δίκτυο με αποτέλεσμα στην επομένη επανάληψη όλοι να ξεκινήσουν απο την ίδια πολιτική [30] [7].



Σχήμα 2.7: *A3C VS A2C [7]*

2.2.3 Trust Region Policy Optimization (TRPO)

Μέχρι στιγμής, οι policy-gradient μέθοδοι υπολογίζουν την πιο απότομη κατεύθυνση ανάβασης για να μεγιστοποιήσουν την αναμενόμενη ανταμοιβή και ωθούν τη στρατηγική προς αυτή την κατεύθυνση. Χρησιμοποιούν δηλαδή την παράγωγο πρώτης τάξης και προσεγγίζουν την επιφάνεια σαν να είναι επίπεδη. Γίνεται εύκολα αντιληπτό ότι αν η επιφάνεια έχει υψηλή καμπυλότητα, τότε θα οδηγηθούμε σε τελείως λανθασμένη κατεύθυνση με πολύ απότομες κινήσεις. Αυτό το πρόβλημα έρχεται να το λύσει ο TRPO αλγόριθμος, ο οποίος μας εξασφαλίζει ότι η στρατηγική δεν θα αλλάξει απότομα, δηλαδή η πολιτική θα ενημερώνεται σε μια περιοχή εμπιστοσύνης (trust region). Ως περιοχή εμπιστοσύνης ορίζεται μια περιοχή βάσει ένος μέγιστου μέγεθους βήματος, στην οποία αναζητάμε το τοπικό μέγιστο της πολιτικής. Επαναλαμβάνοντας αυτή τη διαδικασία, μπορούμε να καταλήξουμε στο ολικό μέγιστο της πολιτικής, που είναι ο στόχος μας.



Σχήμα 2.8: *TRPO [8]*

Γι' αυτόν τον λόγο προστίθεται ένας περιορισμός στο συγκεκριμένο πρόβλημα βελτιστοποίησης 2.4, γνωστός ως KL-divergence. Ο συγκεκριμένος περιορισμός, KL-Divergence, μεταξύ της παλίας και της νέας στρατηγικής πρέπει να είναι μικρότερος απο μια παράμετρο δ, η οποία αποτελεί το μέγεθος της περιοχής και ονομάζεται περιορισμός περιοχής [32]. Συνεπως,το πρόβλημα μεγιστοποίησης της συνάρτησης ανταμοιβής που δινόταν απο την εξίσωση 2.2 μετασχηματίζεται στο ακόλουθο:

$$\hat{g} = \hat{\mathbb{E}}_t\Big[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\hat{A}_t\Big] \longrightarrow \begin{array}{ll} \underset{\theta}{\text{maximize}} & \hat{\mathbb{E}}_t\left[\dfrac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] \\ \text{subject to} & \hat{\mathbb{E}}_t[\mathrm{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]] \leq \delta. \end{array}$$

Σχήμα 2.9: *TRPO πρόβλημα [9]*

2.2.4 Proximal Policy Optimization (PPO)

Ο αλγόριθμος TRPO είναι ένας πολύ καλός αλγόριθμος, όμως πάσχει απο ένα σημαντικό πρόβλημα και αυτό είναι ο KL περιορισμός που εισάγει. Με τον περιορισμό που εισάγει, το συγκεκριμένο πρόβλημα λύνεται αποτελεσματικά με τη χρήση του αλγορίθμου συζευγμένης κλίσης, έπειτα απο την εφαρμογή μιας γραμμικής προσέγγισης στην συνάρτηση ανταμοιβής και μιας τετραγωνικής προσέγγισης στον περιορισμό. Επιπλέον, η εφαρμογή του TRPO δεν αποτελεί μια straight-forward υπόθεση. Γι' αυτό τον λόγο χρησιμοποιείται πολλές φορές ο PPO, ο οποίος επιτυγχάνει την αξιόπιστη απόδοση του TRPO με χρήση βελτιστοποίηση πρώτης τάξης [9].

Σύμφωνα με [9], ορίζεται ο λόγος μεταξύ της παλιάς και της νέας στρατηγικής ως:

$$r_t(\partial) = \frac{\pi_{\partial}(a_t|s_t)}{\pi_{\partial old}(a_t|s_t)} \tag{2.2}$$

Οπότε η αντικειμενική συνάρτηση, δηλαδή η συνάρτησης αναμενόμενης ανταμοιβής, του TRPO γράφεται ως:

$$J(\partial) = \hat{E}_t[r_t(\partial)\hat{A}_t] \tag{2.3}$$

Όμως χωρίς τον περιορισμό της απόστασης (KL περιορισμός) ανάμεσα στην παλία και την καινούργια στρατηγική, η μεγιστοποίηση της παραπάνω συνάρτησης μπορεί να οδηγήσει σε εξαιρετικά μεγάλες ενημερώσεις της πολιτικής. Γι αυτόν τον λόγο, ο PPO αλγόριθμος επιβάλλει τον περιορισμό αναγκάζοντας την $r_t(\partial)$ να παραμείνει σε ένα μικρό διάστημα [1-ε,1+ε], όπου η ε αποτελεί υπερπαράμετρο του προβλήματος. Συνεπώς η αντικειμενική συνάρτηση προς μεγιστοποίηση είναι:

$$J(\partial) = E[min(r_t(\partial)\hat{A}_t, clip(r_t(\partial), 1-\varepsilon, 1+\varepsilon)\hat{A}_t] \tag{2.4}$$

Στην 2.4, ο πρώτος όρος είναι η εξίσωση 2.3 και ο δεύτερος όρος ειναι λόγος μεταξύ της παλιάς και της νέας στρατηγικής στο διάστημα [1-ε,1+ε] και λαμβάνει το ελάχιστο απο αυτούς τους όρους, δηλαδή παίρνει την ελάχιστη τιμή μεταξύ της αρχικής τιμής του $r_t(\partial)$ και της αποκομμένης έκδοσης του. Με αυτόν τον τρόπο δεν γίνεται να υπάρξουν απότομες αλλαγές ανάμεσα στην παλία και την νέα στρατηγική και παράλληλα είναι ένα πρόβλημα βελτιστοποίησης πρώτης τάξης, χωρίς την εισαγωγή περιορισμών [9] [7].

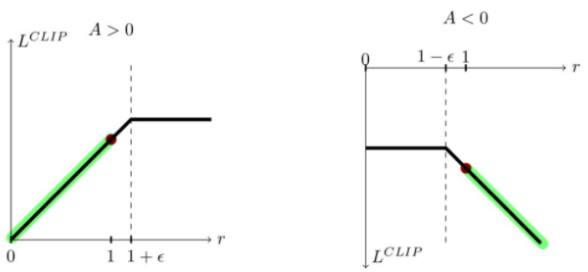


Figure 1: Plots showing one term (i.e., a single timestep) of the surrogate function L^{CLIP} as a function of the probability ratio r, for positive advantages (left) and negative advantages (right). The red circle on each plot shows the starting point for the optimization, i.e., $r = 1$. Note that L^{CLIP} sums many of these terms.

Σχήμα 2.10: *PPO [9]*

Σύμφωνα με την 2.10, το αριστερό γράφημα δείχνει ότι για θετικές τιμές της Advantage συνάρτησης, όταν αυξάνεται η $r_t(\partial)$, δηλαδή η ενέργεια που έχει επιλεχθεί είναι πιο πιθανή σύμφωνα με τη νέα πολιτική απο ό,τι με την παλιά, τότε η αντικειμενική συνάρτηση κόστους ($J(\partial) = L^{CLIP}$) σταθεροποιείται. Το αντίστοιχο συμβαίνει όταν η Advantage συνάρτηση είναι αρνητική και η επιλεχθείσα ενέργεια είναι λιγότερη πιθάνη με τη νέα στρατηγική απο ό,τι με την παλιά. Με αυτόν τον τρόπο, ελέγχεται η ανανέωση της στρατηγικής ώστε να μην αυξάνεται απότομα [33].

Συνοψίζοντας τους παραπάνω 3 αλγορίθμους, στον πινακα παρουσιάζονται οι αντικειμενικές συναρτήσεις που θέλουμε να μεγιστοποιήσουμε:

Πίνακας 2.1: *Αντικειμενική συνάρτηση για κάθε αλγόριθμο*

Αλγόριθμος	Αντικειμενική συνάρτηση
A2C	$E[\log(\pi_\partial(a_t\|s_t)A(t)]$
TRPO	$\hat{E}[r_t(\partial)\hat{A}_t]$
PPO	$E[min(r_t(\partial)\hat{A}_t, clip(r_t(\partial), 1-\varepsilon, 1+\varepsilon)\hat{A}_t]$

Εντροπία

Σύμφωνα με [31], προσθέτοντας την εντροπία της στρατηγικής στην αντικειμενική συνάρτηση, βελτιώθηκε η εξερεύνηση της καλύτερης πολιτικής με αποτελέσμα να αποθαρρύνεται η πρόωρη σύγκλιση σε μη βέλτιστες ντετερμινιστικές πολιτικές [34]. Λαμβάνοντας το gradient για κάθε μία απο την παραπάνω αντικειμενικές συναρτήσεις, εισάγεται ένας νέος όρος $\beta\nabla H(\pi)$, όπου η υπερπαράμετρος β ρυθμίζει την ισχύ του όρου κανονικοποίησης της εντροπίας.

Μέρος II

Πρακτικό Μέρος

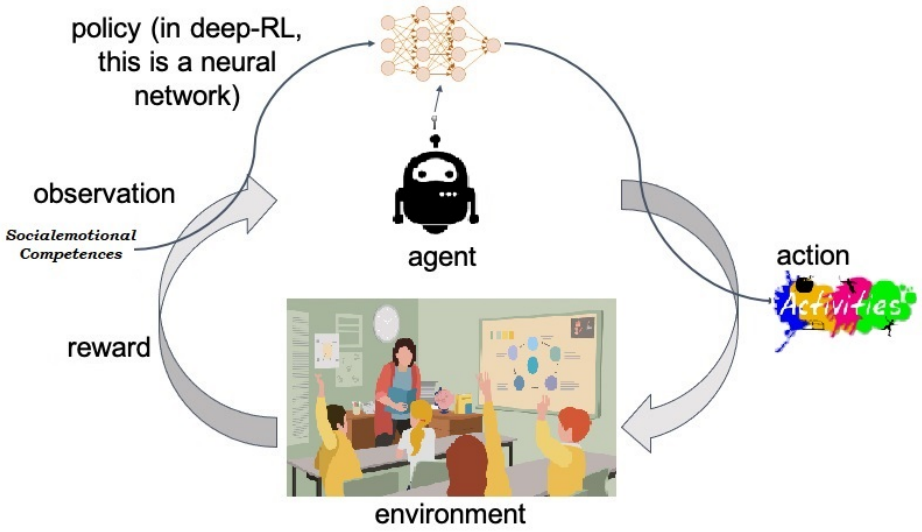
Κεφάλαιο 3

Σχεδίαση του διαδραστικού συστήματος

Στο κεφάλαιο αυτό παρουσιάζεται η μελέτη που έγινε για τη σχεδίαση του διαδραστικού συστήματος σύστασης σε ένα υψηλότερο επίπεδο.

3.1 Ανάλυση του συστήματος

Σύμφωνα με την 3.1, ενας πράκτορας αλληλεπιδρά με ένα σύνολο μαθητών εντός μιας σχολικής τάξης και τους προτείνει συστάσεις για την υλοποίηση εκπαιδευτικών δραστηριοτήτων που έχουν ως στόχο τη βελτίωση των κοινωνικών και συναισθηματικών ικανοτήτων τους μέσα στην ομάδα. Αναλυτικότερα, μια ομάδα απο μαθητές διαθέτει αναπτυγμένες ή μη κοινωνικές και συναισθηματικές ικανότητες, τις οποίες χρειάζεται να βελτιώσει για την αρμονική συνύπαρξη των μαθητών στην τάξη.



Σχήμα 3.1: *Σχεδιασμός του διαδραστικού συστήματος σύστασης*

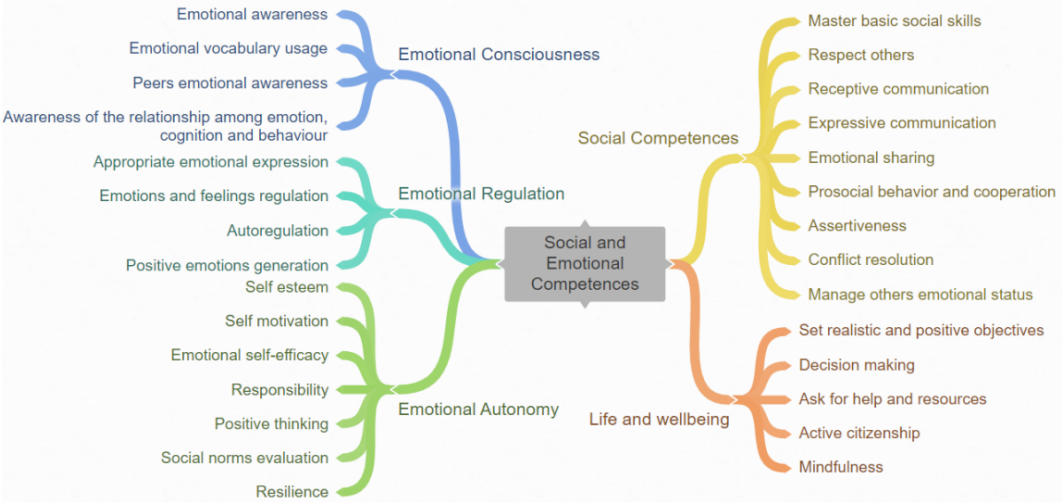
3.1.1 Κοινωνικές και Συναισθηματικές Δραστηριότητες

Κάθε δραστηριότητα χαρακτηρίζεται απο μία λίστα κοινωνικών και συναισθηματικών δεξιοτήτων που μπορεί να βελτιώσει. Σύμφωνα με το μοντέλο που περιγράφεται απο [10], οι

δεξιότητες στις οποίες στοχέυει η δραστηριότητα χωρίζονταισε 5 κατηγορίες, οι οποίες είναι:

- **Συναισθηματική συνείδηση** : Η ικανότητα να έχει κάποιος συναισθηματική επίγνωση ή να έχει εμπλουτισμένο συναισθηματικό λεξιλόγιο για να μπορεί να προσδιορίσει τα συναισθήματα του και να είναι υπεύθυνος των πράξεων του.
- **Συναισθηματική αυτονομία** : Η ικανότητα να απελευθερωθεί κάποιος από τη συναισθηματική εξάρτηση από τους γονείς τους ή άλλους ανθρώπους και να είναι υπεύθυνος για ό,τι του συμβαίνει.
- **Συναισθηματική αυτορύθμιση** : Η ικανότητα να ασκεί κανείς έλεγχο στη συναισθηματική του κατάσταση. Μπορεί να περιλαμβάνει συμπεριφορές όπως η επανεξέταση μιας δύσκολης κατάστασης για τη μείωση του θυμού ή άγχους, η απόκρυψη ορατών σημαδιών λύπης ή φόβου ή η εστίαση σε λόγους με σκοπό τη χαρά και την ηρεμία.
- **Ζωή και WellBeing**: Οι ικανότητες που σχετίζονται με το να είναι κάποιος ενεργός πολίτης ή να θέτει ρεαλιστικούς στόχους και να μπορεί να πάρει αποφράσεις.
- **Κοινωνικές ικανότητες**: Η κοινωνική ικανότητα αποτελείται από κοινωνικές, συναισθηματικές, γνωστικές και συμπεριφορικές δεξιότητες που απαιτούνται για την επιτυχή κοινωνική προσαρμογή.

Αυτές οι παραπάνω 5 κατηγορίες διασπώντονται σε 30 ικανοτήτες μικρότερης κλίμακας, που χρησιμοποιούνται για τον χαρακτηρισμό των δραστηριοτήτων. Οπότε, κάθε δραστηριότητα μπορεί να αφορά περισσότερες από μία ικανότητες μικρότερης κλίμας προς βελτίωση για μια χρονική στιγμή. Συνεπώς, η αναπαράσταση κάθε δραστηριότητας ειναι ένα διανύσμα μήκους 30, το οποίο λαμβάνει τιμές 0 ή 1 για το αν απευθύνεται στην συγκεκριμένη δεξιότητα μικρής κλίμακας.



Σχήμα 3.2: *Κοινωνικές και Κοινωνικές Ικανότητες [10]*

Πέραν των ικανοτήτων στις οποιές αναφέρεται κάθε δραστηριότητα, χαρακτηρίζεται και απο τη διαρκεία της. Η διάρκεια μιας δραστηριότητας είναι καθοριστικής σημασίας για το πρόβλημα μας, αφού ένα σύνολο δεξιότητων μιας ομάδας πρέπει να βελτιωθεί εντός ενος

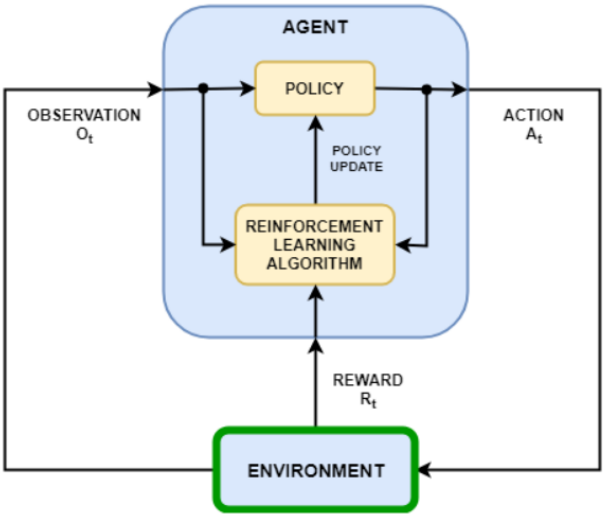
συγκεκριμένου χρονικού προϋπολογισμού. Αυτό θέτει ιδιαίτερο ενδιαφέρον στο σύστημα σύστασης, διότι του θέτει ένα χρονικό εύρος στο οποίο μπορεί να συστήνει δραστηριότητες με σκοπό τη βελτίωση των ικανοτήτων της ομάδας, με αποτελέσμα να εισέρχεται ο παράγοντας της ρεαλιστικότητας με παράλληλο στόχο τη βελτίωση της ομάδας σε αυτό το χρονικό περιθώριο.

3.1.2 Κατάσταση του περιβάλλοντος

Το περιβάλλον ενός προβλήματος ενισχυτικής μάθησης χαρακτηρίζεται απο την κατάσταση στην οποιά βρίσκεται. Στο δικό μας πρόβλημα, η κοινωνική και συναισθηματική κατάσταση μιας εκπαιδευτικής ομάδας χαρακτηρίζεται απο τις δεξιότητες μικρές κλίμακας που αναφέρονται στο 3.1.1. Κάθε ικανότητα μικρή κλίμακας λαμβάνει τιμές απο το -1 εώς το 1, το οποίο υποδηλώνει αν είναι καλα αναπτυγμένη ή όχι. Στόχος του συστήματος σύστασης είναι να προτείνει δραστηριότητες που ταιριάζουν με τις κοινωνικές και συναισθηματικές ανάγκες της ομάδας. Όταν μια ικανότητα μικρής κλίμακας έχει ανάγκη να βελτιωθεί, τότε η τιμή της είναι πιο κοντά στο -1 και τότε πρέπει να εφαρμοστει η κατάλληλη ενέργεια η οποία να λαμβάνει την τιμή 1 για αυτή την ικανότητα. Αντίστοιχα, όταν μια ικανότητα μικρής κλίμακας της ομάδας έχει τιμή κοντά στο 1, τότε δεν χρειάζεται να εφαρμοστεί κάποια ενέργεια που να στοχεύει σε αυτήν συγκεκριμένα. Συνολικά, οι υψηλές τιμές μιας ικανότητας μικρής κλίμακας μιας εκπαιδευτικής ομάδας αντικατοπτρίζουν την ανάγκη βελτίωση της με διεξαγωγή δραστηριότήτων στοχευμένων σε αυτή και που παράλληλα έχουν τη χρονική διάρκεια που τους επιτρέπει να τη βελτιώσουν.

Συνδυάζοντας λοιπόν τα παραπάνω, μια εκπαιδευτική ομάδα χαρακτηρίζεται απο ένα διάνυσμα 30 θέσεων με τιμες στο διάστημα [-1,1],όπου κάθε τιμή δηλώνει την ικανότητα μικρής κλίμακας και προτείνονται σε αυτή την ομάδα δραστηριότητες που αναπαριστώνται απο ένα διάνυσμα 30 θέσεων με τιμές 0 ή 1 και παράλληλα κάθε δραστηριότητα που προτείνεται έχει μια χρονική διάρκεια υλοποίησης, με αποτέλεσμα σε κάθε χρονική στιγμή να μείωνεται ο χρονικός προυπολογισμός που έχει η ομάδα για να βελτιωθεί.

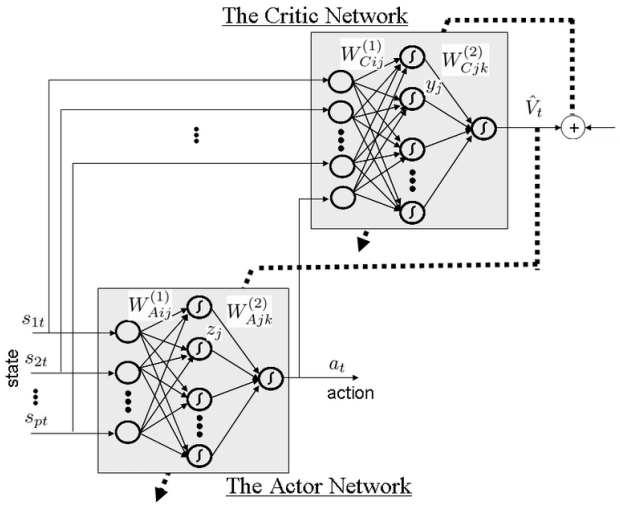
Σύμφωνα με το σχήμα 3.4, έχει γίνει η περιγραφή του περιβάλλοντος, των παρατηρήσεων και των ενεργειών που λαμβάνονται. Όπως φαίνεται, το σχήμα αποτελείται επίσης απο τον πράκτορα ενισχυτικής μάθησης, ο οποίος αποτελείται απο την πολιτική προς βελτιστοποίηση και τον αλγόριθμο ενισχυτικής μάθησης.



Σχήμα 3.3: *Σχεδιασμός συστήματος απο την οπτική της ενισχυτικής μάθησης[11]*

3.1.3 Πολιτική του προβλήματος ενισχυτικής μάθησης

Όπως γνωρίζουμε,η πολιτική ή αλλιώς στρατηγική σε ένα πρόβλημα ενισχυτικής μάθησης, είναι μια συνάρτηση που λάμβανει σαν είσοδο την κατάσταση s και δινει σαν έξοδο την ενέργεια a. Με αυτόν τον τρόπο, η πολιτική χρησιμοποιείται απο τον πράκτορα για να αποφασίσει ποια ενέργεια a να πραγματοποιήσει όταν βρίσκεται σε μια συγκεκριμένη κατάσταση s. Στο συγκεκριμένο πρόβλημα σύστασης, η πολιτική αναπαριστάται μέσω νευρωνικού δικτύου με τη λογική που δουλεύει ο Actor to Critic. Ο Actor αναπαρίσταται απο ένα νευρωνικό δίκτυο με δύο στρώματα, το οποίο παράγει την καλύτερη ενέργεια για μια συγκεκριμένη κατάσταση και ο Critic απο την άλλη πλευρά αναπαρίσταται επίσης ως ένα νευρωνικό δίκτυο που λαμβάνει τις καταστάσεις απο το περιβάλλον και την ενέργεια απο τον Actor και δίνει σαν έξοδο την Value συνάρτηση για τον συγκεκριμένο συνδυασμό.



Σχήμα 3.4: *Πολιτική του συστήματος σύστασης [12]*

3.1.4 Αλγόριθμος ενισχυτικής μάθησης

Τέλος, ο πράκτορας απαρτίζεται απο τον αλγόριθμο ενισχυτικής μάθησης και την πολιτική. Σχετικά με την πολιτική ενισχυτικής μάθησης, χρησιμοποιείται η λογική του Actor to Critic που αναφέρεται παραπάνω. Όσον αφορά τον αλγόριθμο, στην 2.2.1 ανάφερονται με περισσότερες λεπτομέρειες οι αλγόριθμοι A2C,PPO και TRPO. Η λογική που ακολουθείται είναι ότι ο Critic εκτιμά την Value συνάρτηση και ο Actor χρησιμοποιεί ένα policy gradient αλγόριθμο για να παράξει την καλύτερη ενέργεια.

Στο 4.2 παρουσιάζεται με περισσότερες τεχνικές λεπτομέρειες η υλοποίηση του προβλήματος με τη χρήση της βιβλιοθήκης Stable-Baselines.

Κεφάλαιο 4

Υλοποίηση

Στο κεφάλαιο αυτό περιγράφεται η υλοποίηση του συστήματος, με βάση τη μελέτη που παρουσιάστηκε στο προηγούμενο κεφάλαιο. Αρχικά παρουσιάζονται τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν. Στη συνέχεια δίνονται οι λεπτομέρειες υλοποίησης με τα συγκεκριμένα εργαλεία.

4.1 Σύγκριση των εργαλείων

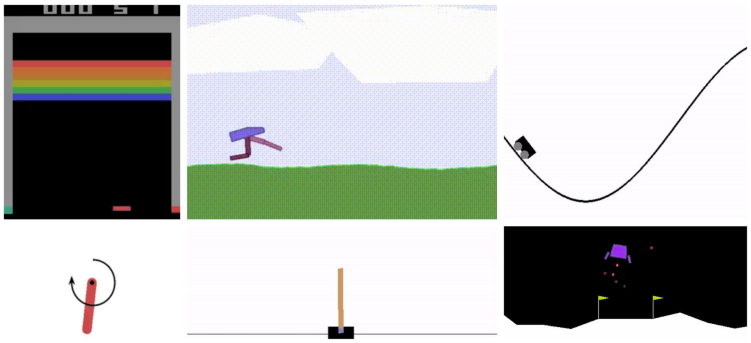
Στη συγκεκριμένη ενότητα παρουσιάζεται η σύγκριση μεταξύ των διαφορετικών frameworks και βιβλιοθηκών. Το συγκεκριμένο σύστημα σύστασης υλοποιήθηκε με τη βοήθεια της προγραμματιστικής γλώσσας *Python*.

Στο συγκεκριμένο τομέα, δηλαδή στην υλοποίηση διαδραστικών συστημάτων, έχουν γίνει διάφορες προσεγγίσεις, οι οποίες είναι πολύ στοχευμένες στο πρόβλημα που υλοποιούν, πχ. μπορεί να αφορούν τη δημιουργία ενός διαδραστικού συστήματος για τη σύσταση προιόντων στις διαδικτυακές διαφημίσεις. Όπως γίνεται αντιληπτό, το γεγονός ότι αυτά τα συστήματα είναι υλοποιήμενα για πολύ συγκεκριμένους τομείς κάνει το πρόβλημα μας πιο δύσκολο προς υλοποίηση αλλα παράλληλα τρομερά ενδιαφέρον. Παραδείγματα τέτοιων συστημάτων είναι το Gym-RecSys και το Reco-Gym που αφορουν τον τομέα της σύστασης ταινιών και προιόντων στις διαδικτυακές διαφημίσεις με βοήθεια ενισχυτικής μάθησης αντίστοιχα. Η πλειοψήφια τέτοιων συστημάτων κάνει χρήση της εργαλειοθήκης Open AI Gym και των διαφορετικών frameworks, όπως το Tensorflow ή το Pytorch. Όσον αφορά το δικό μας πρόβλημα χρησιμοποιείται για τη δημιουργία του περιβάλλοντος η εργαλειόθηκη Open AI Gym και η βιβλιοθήκη Stable-Baselines για την εφαρμογή αλγορίθμων ενισχυτικής μάθησης, ύστερα απο σύγκριση της με παρόμοιες βιβλιοθήκες.

4.1.1 Open AI Gym

Η πλειοψηφία αυτών των προβλημάτων υλοποείται με τη χρήση του Open AI Gym.Το Open AI Gym είναι μια εργαλειοθήκη για την ανάπτυξη και τη σύγκριση αλγορίθμων ενισχυτικής μάθησης. Αποτελεί μια διεπαφή ανοιχτού κώδικα για πρόβληματα ενισχυτικής μάθησης και παρέχει ένα περιβάλλον πάνω στο οποίο ο προγραμματιστής μπορεί να εφαρμόσει αλγόριθμους ενισχυτικής μάθησης. Είναι πολύ εύκολη στη χρήση και συμβατή με βιβλιοθήκες αριθμητικών υπολογισμών, όπως το Tensorflow και το Theano. Η συγκεκρι-

μένη εργαλειοθήκη είναι ευρώς γνωστή στον τομέα της ενισχυτικής μάθησης. Το γεγονός ότι περιέχει μια συλλογή περιβάλλοντων(προβλημάτων) όπως παιχνίδια πινγκ-πονγκ ή ρομποτ που περπατάνε, τα οποία μπορούν να λύσουν οι εκπαιδευτικοί πράκτορες την κάνει τόσο διάσημη τη συγκεκριμένη βιβλιοθήκη.



Σχήμα 4.1: *Συστήματα υλοποιημένα με Open AI Gym*

4.1.2 Βιβλιοθήκες αλγορίθμων ενισχυτικής μάθησης

Βασισμένα σε αυτή την εργαλειοθήκη, τα τελευταία χρόνια έχουν αναπτυχθεί πολλές βιβλιοθήκες ενισχυτικής μάθησης με σκοπό την παροχή των απαραίτητων εργαλειών τόσο για την εφαρμογή όσο και για τη δοκιμή αλγόριθμων μηχανικής μάθησης. Παρόλαυτά διαφέρουν αρκετά και ήταν σκόπιμο να γίνει η επιλογή της κατάλληλης βιβλιοθήκης, η οποία θα είναι γρήγορη,αξιόπιστη και σχετική το συγκεκριμένο πρόβλημα προς υλοποίηση. Ορισμένές τέτοιες βιβλιοθήκες παρουσιάζονται στον ακόλουθο πίνακα.

Πίνακας 4.1: *Βιβλιοθήκες Ενισχυτικής μάθησης*

Βιβλιοθήκη	**Περιγραφή**
Tensorforce	Tensorforce είναι μια βιβλιοθήκη Deep RL ανοιχτού κώδικα που βασίζεται στο framework Tensorflow της Google
TF-Agents	TFAgents είναι μια βιβλιοθήκη της Python που έχει σχεδιαστεί για να διευκολύνει την υλοποίηση, την ανάπτυξη και τη δοκιμή αλγορίθμων RL
Stable-Baselines	Stable Baselines είναι ένα σύνολο βελτιωμένων υλοποιήσεων των αλγορίθμων RL που βασίζονται στο OpenAI Baselines

4.1.3 Stable-Baselines

Αυτή που ξεχώρισε είναι η Stable-Baselines, η οποία παρέχει πληθώρα απο State of the art αλγορίθμους, όπως ο A2C,TRPO,SAC κλπ. και είναι συμβατή με το Tensorflow και πιο συγκεκριμένα με το Tensorboard, το οποίο είναι ενα σετ εργαλείων του Tensorflow για οπτικοποίηση πειραμάτων μηχανικής μάθησης. Η συμβατότητα της συγκεκριμένης βιβλιοθήκης με το Tensorboard δίνει δυνατότητες όπως η παρακολούθηση και η οπτικοποίηση των μετρικών κατα τη διαδικασία εκπαίδευσης του μοντέλου μηχανικής μάθησης ή τη σύγκριση μετρικών ανάμεσα σε διαφορετικούς αλγορίθμους. Η συγκεκριμένη βιβλιοθήκη δεν περιορίζεται μόνο σε αυτα τα σημαντικά πλεονεκτήματα που παρέχει αλλά προσφέρει

και ένα εξαιρετικό documentation, ακόμη και για τη δημιουργία custom περιβάλλοντος και πληθώρα tutorials και παραδειγμάτων.

4.2 Λεπτομέρειες υλοποίησης

Αρχικά δημιουργείται ένα custom περιβαλλον που υλοποιεί το πρόβλημα μας. Σύμφωνα με τη βιβλιοθήκη Stable-Baselines, για τη δημιουργία ενός προσαρμοσμένου περιβάλλοντος χρείαζεται το περιβάλλον να κληρονομεί απο την OpenAI Gym κλάση. Με αυτόν τον τρόπο το περιβάλλον που δημιουργούμε ακολουθεί τη διεπαφή του OpenAI Gym και στη συνέχεια χρησιμοποιούνται οι αλγόριθμοι της Stable-Baselines.

4.2.1 OpenAI Gym περιβάλλον

Ένα περιβάλλον που ακολουθεί την OpenAI Gym διεπαφή, απαρτίζεται απο 3 κύριες μεθόδους:

- **Reset()** : Χρησιμοποιείται στην αρχή ενός επεισοδίου και επιστρέφει μια παρατήρηση του προβλήματος.

- **Step(action)**: Λαμβάνει μια ενέργεια απο το περιβάλλον και επιστρέφει την επόμενη παρατήρηση, την άμεση ανταμοιβή, αν το επεισόδιο έχει τελειώσει και πρόσθετες πληροφορίες που καθορίζονται απο τον χρήστη.

- **Render()** : Είναι μια προαιρετική μέθοδος και χρησιμοποιείται αν θέλουμε να οπτικοποιήσουμε τον πράκτορα στην πράξη. Χρησιμοποιείται με έτοιμα περιβάλλοντα απο το OpenAI Gym.Στο δικό μας πρόβλημα δεν εφαρμόζεται.

Επιπρόσθετα, κάθε περιβάλλον που ακολουθεί τη διεπαφή του OpenAI Gym, αποτελείται απο τον χώρο των παρατηρήσεων και τον χώρο των ενεργειών, όπως παρουσιάζονται ακολούθως:

- **Observation space** : Στο δικό μας πρόβλημα είναι ένα box, δηλαδή ένας τύπος δομών δεδομένων που ονομάζονται Spaces και παρέχονται απο το OpenAI Gym. Πιο συγκεκριμένα, ο χώρος των παρατηρήσεων είναι ένα μονοδιάστατο διάνυσμα 30 θέσεων με τιμές απο -1 εως 1.

- **Action space** : Στο δικό μας πρόβλημα είναι ένα MultiBinary, δηλαδή ένας τύπος δομών δεδομένων που ονομάζονται Spaces και παρέχονται απο το OpenAI Gym. Στο δικό μας πρόβλημα, ο χώρος των ενεργειών είναι ένα μονοδιάστατο διάνυσμα 30 θέσεων με τιμές απο 0 ή 1.

Τέλος, έχοντας υλοποίησει το περιβάλλον, η βιβλιοθήκη Stable-Baselines παρέχει βοηθητική συνάρτηση,η οποία ελέγχει αν το περιβάλλον ακολουθεί την Open AI gym διεπαφή και ελέγχει αν ειναι συμβατό με την ίδια, για την εφαρμογή των αλγορίθμων.

4.2.2 Δημιουργία custom OpenAI Gym περιβάλλοντος

Σύμφωνα με όσα αναφέρονται στην ενότητα 3.1, ορίζονται τα παρακάτω για το περιβάλλον μας:

- **State** : Numpy διάνυσμα μήκους 30 με τιμές στο διάστημα [-1,1]
- **Time_Budget** : Δηλώνει τον χρονικό προυπολογισμό για την εκπαιδευτική ομάδα. Σε κάθε αρχή επεισοδίου λαμβάνει την τιμή 3600 και η μονάδα μέτρησης του ειναι τα λεπτά.
- **Activity_Duration** : Δηλώνει τη χρονική διάρκεια κάθε δραστηριότητας και λαμβάνει τιμές απο 30 λεπτά ώς 120 λεπτά, δηλαδη μια δραστηριότητα μπορει να διαρκέσει απο μισή ώρα έως 2 ώρες. Σε κάθε γύρο ο χρονικός προυπολογισμός μειώνεται απο τη χρονική διάρκεια της δραστηριότητας που προτάθηκε απο τον πράκτορα.

Μέσω της συνάρτησης step ο πράκτορας λαμβάνει μια ενέργεια απο το περιβάλλον και την αξιολογεί επιστρέφοντας την επομένη παρατήρηση, την ανταμοιβή και το αν έχει τελειώσει ή όχι το επεισοδίο. Συνεπώς, ο τρόπος αξιολόγησης καθορίζεται απο τη μοντελοποίηση της ανταμοιβής, που περιγράφεται στην επομένη ενότητα.

Ανανέωση της κατάστασης και ανταμοιβή του πράκτορα

Σε κάθε χρονικό βήμα που πραγματοποιείται υπολογίζεται το ενδιαφέρον της εκπαιδευτικής ομάδας σχετικά με την ενέργεια που προτάθηκε. Το συγκεκριμένο ενδιαφέρον δίνεται απο τον τύπο:

$$Group\_interest = action \cdot state \tag{4.1}$$

Αυτός ο τύπος μας δείχνει οτι αν η δραστηριότητα που προτείνεται έχει τιμές 1 για τις ικανότητες μικρής κλίμακας που έχουν ανάγκη για βελτίωση, δηλαδή τιμές κοντά στο 1, τότε θα λαμβάνει υψηλότερες τιμές. Στην αντίθετη περίπτωση, που προτάθηκε μια δραστηριότητα, η οποία απευθύνεται σε ικανότητες μικρής κλίμακας με τιμές κοντά στο -1, τότε το Group_interest θα λαμβάνει χαμηλότερες τιμές. Σύμφωνα με το ενδιάφερον της εκπαιδευτικής ομάδας για την ενέργεια, υπολογίζεται η μεταβλητή positive_update_probability, που δηλώνει αν η ενέργεια που προτάθηκε καλύπτει τις ανάγκες της ομάδας, λαμβάνοντας τιμές στο διάστημα [0,1].

$$Positive\_update\_probability = [Group\_interest + sum(action)]/(2sum(action)) \tag{4.2}$$

Στην παράπάνω σχέση, αθροίζοντας τις τιμές του διανύσματος της ενέργειας προκύπτει ο αριθμός των ικανοτήτων μικρής κλίμακας, στις οποίες απευθύνεται και με αυτόν τον τρόπο πραγματοποιείται η κανονικοποίηση. Λαμβάνοντας υπόψιν τη συκγκεκριμένη πιθανότητα, πραγματοποιείται η ανανέωση της κατάστασης και η ανταμοιβή του πράκτορα.

Στη συνέχεια, χρησιμοποιώντας την positive_update_probability πραγματοποιείται η ανανέωση της κατάστασης, με τον ακόλουθο τρόπο:

- **positive_update_probability $\geqslant$ 0.45** : Αν η συγκεκριμένη πιθάνοτητα είναι μεγάλη, τότε η ενέργεια που προτείνεται στοχεύει σχεδόν σε όλες τις ικανότητες μικρής κλίμακας, που έχουν ανάγκη για βελτίωση. Όπως ανάφερεται και παραπάνω, όσο μεγαλύτερες τιμές λαμβάνει αυτή η πιθανότητα τόσο πιο στοχευμένη είναι η δραστηριότητα που προτείνεται στην εκπαιδευτική ομάδα. Σε αυτή την περίπτωση, οι ικανότητες μικρής κλίμακας που έχουν ανάγκη για βελτίωση και στις οποίες απευθύνεται η δραστηριότητα μειώνονται κατα 30%.

- **0.2 $\leqslant$ positive_update_probability $<$ 0.45** : Αν η συγκεκριμένη πιθάνοτητα δεν είναι πολύ μεγάλη, δηλαδή η ενέργεια απευθύνεται και σε δραστηριότητες που δεν έχουν ανάγκη για βελτίωση, είναι δηλαδή καλά αναπτυγμένες, τότε ορίζεται επιπλέον και μια ποινή. Εκείνες οι ικανότητες μικρής κλίμακας της εκπαιδευτικής ομάδας που έχουν ανάγκη για βελτίωση μειώνονται κατα 20% ενώ εκείνες που είναι καλά αναπτυγμένες αυξάνονται κατα 10 %. Αυξάνονται ή μειώνονται μόνο εκείνες οι ικανότητες μικρής κλίμακας, στις οποίες στοχεύει η ενέργεια που συστήνεται.

- **(positive_update_probability $<$ 0.2) & (% of well developed competences $>$ 0.8)** : Σε αυτή την περίπτωση, δηλαδή που η πιθανότητα είναι πολύ χαμηλή και παράλληλα το ποσοστό των καλά αναπτυγμένων ικανότητων μικρής κλίμακας της ομάδας είναι μεγαλύτερο απο 80 %, τότε τερματίζεται νωρίτερα το επεισόδιο. Εφοσόν σχεδόν όλες οι ικανότητες μικρής κλίμακας είναι στο επιθυμητό επίπεδο, δηλαδή δεν θέλουν περαιτέρω βελτίωση και παράλληλα η ενέργεια απευθύνεται σε κάποιες απο αυτές, τότε είναι λογικό η πιθανότητα να λαμβάνει χαμηλές τιμές και να σταμάταει νωρίτερα το επεισόδιο.

- **(positive_update_probability $<$ 0.2)** : Αυτή η περίπτωση δηλώνει την ποινή για τον πράκτορα, δηλαδή αν η ενέργεια που συστήνεται δεν στοχεύει στις ικανότητες μικρής κλίμακας που έχουν ανάγκη και υπάρχουν άλλες που χρειάζονται τότε αυξάνονται κατα 20 %.

Στις παραπάνω περιπτώσεις, ως καλά αναπτυγμένες ικανότητες μικρής κλίμακας ορίζονται εκείνες με αρνητικές τιμές ενώ ως ικανότητες μικρής κλίμακας προς βελτίωση ορίζονται εκείνες με θετικές ή μηδενικές τιμές. Δεδομένου ότι η κατάσταση του περιβάλλοντος λαμβάνει τιμές στο [-1,1], τότε για να πραγματοποιηθεί ποσοστιαία μεταβολή στις παραπάνω περιπτώσεις, κανονικοποιείται στο [0,1] και μολις υπολογιστεί η μεταβολή επανέρχεται στο αρχικό της εύρος τιμών.

Τέλος, η ανταμοιβή του πράκτορα υπολογίζεται με τον ακόλουθο τρόπο: Ελέγχεται αν υπάρχουν ικανότητες μικρής κλίμακας, στις οποίες δεν απευθύνεται η δραστηριότητα και έχουν μεγάλη ανάγκη για βελτίωση, έχουν δηλαδη μεγαλύτερες τιμές απο τη λιγότερο ανεπτυγμένη ικανότητα, στην οποία εφαρμόζεται η ενέργεια.

- **Χωρίς ποινή** : Αν δεν ύπαρχουν τέτοιες ικανότητες, στις οποίες δεν έχει εφαρμοστεί κάποια είδους ενέργεια, τότε η ανταμοιβή του πράκτορα είναι το ποσοστό των καλά αναπτυγμένων δραστηριότητων (που τους εφαρμόστηκε η ενέργεια).

- **Με ποινή** : Αν όμως τέτοιες ικανότητες υπάρχουν, τότε η ανταμοιβή του πράκτορα δέχεται μια ποινή και ορίζεται ως η διαφορά του ποσοστού των καλά αναπτυγμένων δραστηριοτήτων (που τους εφαρμόστηκε η ενέργεια) με το ποσοστό αυτών των ικανοτήτων.

Το επεισόδιο τερματίζει είτε όταν δεν υπάρχει περαιτέρω χρονικός προυπολογισμός είτε όταν αριθμός των καλά αναπτυγμένων ικανοτήτων μικρής κλίμακας ειναι μεγαλύτερος του 27 (δηλαδή το 80%) είτε όταν τερματίζεται νωρίς λογώ μη απαίτησης περαιτέρω βελτίωσης.

Παρακάτω παρουσιάζεται η υλοποιήση που περιγράφεται, στην προγραμματισική γλώσσα Python.

```
import numpy as np
import gym
from gym import spaces
import random
import matplotlib.pyplot as plt
import tensorflow as tf

class EduRecEnv(gym.Env):
  """
  Custom Environment that follows gym interface.
  """
  metadata = {'render.modes': ['console']}

  def __init__(self):
    super(EduRecEnv, self).__init__()



    # Define action and observation space
    # They must be gym.spaces objects

    # Action space
    self.action_space = spaces.MultiBinary(30)

    # Observation space
    self.observation_space = spaces.Box(low=-1, high=1,
                                        shape=(30,), dtype=np.float64)
    # State
    np.random.seed(1)
    self._state = np.random.uniform(-1,1,30)

    # Time budget (Minutes)
    self._time_budget = 3600
```

```
def reset(self):

  # Reset state
  np.random.seed(1)
  self._state =np.random.uniform(-1,1,30)
  self._time_budget = 3600
  early_stop = False
  return self._state

def step(self, action):

  # Duration of activity
  self._activity_duration = random.uniform(30.0,120.0)

  # Calculate group's interest given the action
  group_interest = (np.dot(action,self._state))

  # Normalize group's interest in [0,1]
  positive_update_prob = (group_interest+sum(action))/(2*sum(action))

  # Bad & Well developed microcompetences
  condition_bd = self._state >= 0
  condition_wd = self._state < 0

  # Initialize early stopping
  early_stop = False
  # Count the percentage of well developed microcompetences before the update of the
   state
  all_comp = (np.count_nonzero(self._state<=0))/30

  # Normalize state in [0,1]
  self._state = (self._state + 1)/2

  # High values of positive_update_probability indicate that the the proposed action
   tackle the group's social and emotional needs
  # and as a result the micro-competences of the educational group are reduced by
  30% (high shift & no penalty).
  if positive_update_prob >= 0.45:
      # only reward
      self._state[np.where((action==1) & (condition_bd)==True)] = self._state[np.
  where((action==1) &
                                                                                (
```

```
    condition_bd)==True)]*0.7
    # Low values of positive_update_probability reflect that the action proposed by
    the agent is not aimed at the micro-competences of the group
    # that need to be strengthen and so the well-developed micro-competences are
    increased by 20% and the bad-developed are reduced by 10%(lower shift).
    elif (positive_update_prob >=0.2) & (positive_update_prob <0.45):
        # reward
        self._state[np.where((action==1) & (condition_bd)==True)] = self._state[np.
    where((action==1) &

      (condition_bd)==True)]*0.8
        # penalty
        self._state[np.where((action==1) & (condition_wd)==True)] = self._state[np.
    where((action==1)

      & (condition_wd)==True)]*1.1
    # In this case, the positive_update_probability is vey low and simultaneously the
    percentage of well developed micro-competences is greater than 80%.
    # Since almost all micro-competences are at the desired level, i.e they do not
    need further improvement  then it is reasonable
    # the positive_update_probability of receiving low values and perfoming early
    stopping
    elif (positive_update_prob <0.2) & (all_comp >0.8):
        early_stop = True
    # If the proposed action does not tackle the group's social and emotional needs
    and other micro-competences needs further improvement,
    # then they decreased by 20%
    elif (positive_update_prob<0.2):
        self._state[np.where((action==1) & (condition_wd)==True)] = self._state[np.
    where((action==1)

      & (condition_wd)==True)]*1.2
    # Normalize again the state in [-1,1]
    self._state = 2*(self._state) - 1
    # Limit the values of microcompetences in the range [-1,1]
    self._state = np.clip(self._state,-1.0,1.0)


    # If there is at least one microcompetence, that needs imporevement and action
    does not refer to it, then we count the number of these
    # microcompetences, in order to decrease the reward.
    max_bad_comp = max(self._state[np.where(action==1)])
    check = any(i >= max_bad_comp for i in self._state[np.where(action==0)])
```

```
    # Count bad developed microcompetences, where action is not applicable.
    bad_dev = (self._state[np.where(action==0)]>= max_bad_comp).sum()

    # Compute the number of microcompetences  with negative values(i.e. well developed
    )
    well_dev = np.count_nonzero(self._state<=0)
    if (check == True):
        reward = well_dev/30 - bad_dev/30
    else:
        reward = well_dev/30

    # Reduce time budget at each time step
    self._time_budget -= self._activity_duration

    # done : A boolean value stating whether  i t s  time to reset the environment
    again.
    done = bool((self._time_budget<0) | ((self._time_budget<0) & (well_dev>27)) |
                ((self._time_budget>0) & (well_dev>27))|(early_stop==True))

    # Optionally we can pass additional info.
    info = {'activity_duration':self._activity_duration,'update prob':
    positive_update_prob,
            'well dev comp':well_dev,'time budget':self._time_budget,'early_stop':
    early_stop}

    return self._state, reward, done, info

  def render(self, mode='console'):
    if mode != 'console':
      raise NotImplementedError()

  def close(self):
    pass
```

Κεφάλαιο 5

Αποτελέσματα

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα απο την εφαρμογή των αλγορίθμων Advantage Actor to Critic, Trust Region Policy Optimization και Proximal Policy Optimization με το fine tuning των υπερπαραμέτρων τους. Παρουσιάζεται και η σύγκριση μεταξύ τους για την εύρεση του καλύτερου αλγορίθμου.

5.1 Μεθοδολογία

Σε αυτή την ενότητα, εφαρμόζεται καθε αλγόριθμος για το περιβάλλον μας και στη συνέχεια αξιολογείται η απόδοση του. Όπως έχει ήδη αναφερθεί, η βιβλιοθήκη Stable Baselines, παρέχει τη χρήση του Tensorboard. Οπότε για κάθε αλγόριθμο που χρησιμοποιείται, αξιολογείται η εκπαίδευση του στο Tensoboard. Τα βήματα για κάθε αλγόριθμο που ακολουθούνται είναι τα ακόλουθα:

- **Βελτιστοποίηση βασικής παραμέτρου κάθε αλγορίθμου**

- **Εύρεση του συντελεστή της εντροπίας**:Όταν ο πράκτορας μαθαίνει μια πολιτική και μια ενέργεια της επιστρέφει μια θετική ανταμοιβή για μια κατάσταση, τότε μπορεί ο πράκτορας να χρησιμοποιεί πάντα αυτή την ενέργεια στο μέλλον, επειδή γνωρίζει ότι επιστρέφει μια θετική ανταμοιβή. Όμως, μπορεί να υπάρχει μια διαφορετική ενέργεια που να οδηγεί σε υψηλότερη ανταμοιβή, αλλά ο πράκτορας να μην τη δοκιμάσει ποτέ και να εκμεταλλεύεται συνεχώς αυτήν που ήδη γνωρίζει. Με άλλα λόγια, ο πράκτορας μπορεί να "κολλήσει" σε ένα τοπικό μέγιστο, επειδή δεν *εξερευνά* τις υπόλοιπες ενέργειες και κατέπέκταση δεν βρίσκει το ολικό μέγιστο. Με την εντροπία, μπορούμε να ενθαρρυνουμε τον πράκτορα να εξερευνήσει τον χώρο των ενεργειών και να αποφύγει τοπικά μέγιστα της πολιτικής.

- **Σύγκριση αλγορίθμων με άυξηση των βημάτων εκπαίδευσης**

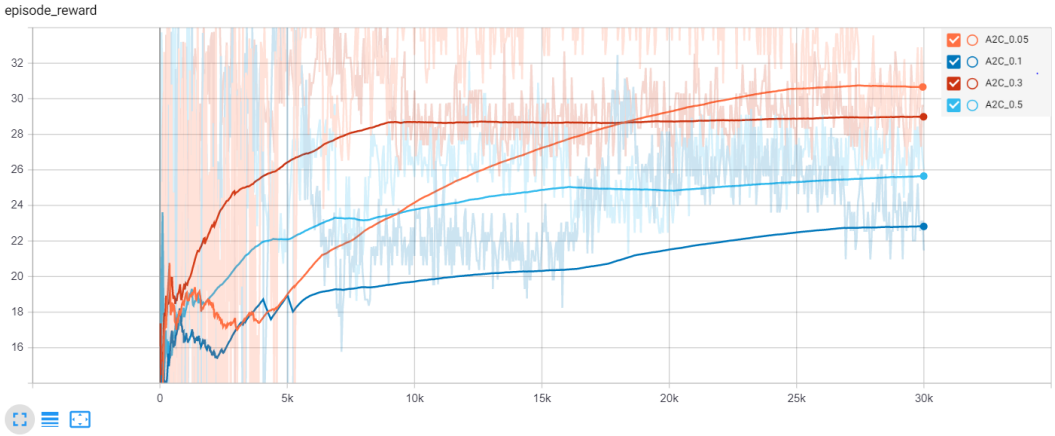
5.2 Αναλυτική παρουσίαση αποτελεσμάτων

5.2.1 Advantage Actor to Critic

Στο συγκεκριμένο αλγόριθμο η ανανέωση των παραμέτρων ∂ της πολιτικής δίνονται απο τον τύπο:

$$\partial = \partial + a\nabla log\pi_{\partial}(a_t|s_t)A(s_t, a_t) \tag{5.1}$$

Η παράμετρος α, η οποία ονομάζεται ρυθμός εκμάθησης, καθορίζει το μέγεθος του βήματος σε κάθε επανάληψη καθώς κινούμαστε προς τη μέγιστη τιμή της συνάρτησης ανταμοιβής. Εκπαιδεύτοντας τον αλγόριθμο Advantage Actor to Critic για 30.000 βήματα και για τιμές του ρυθμού εκμάθησης στο δίαστημα [0.05,0.1,0.3,0.5], φαίνεται απο το διάγραμμα 5.1 ότι τα καλύτερα αποτελέσματα δίνονται για ρυθμό εκμάθησης ίσο με 0.05. Επιπρόσθετα, η ανταμοιβή του πράκτορα με την αύξηση των βημάτων παρουσιάζει σταδιακή αύξηση.

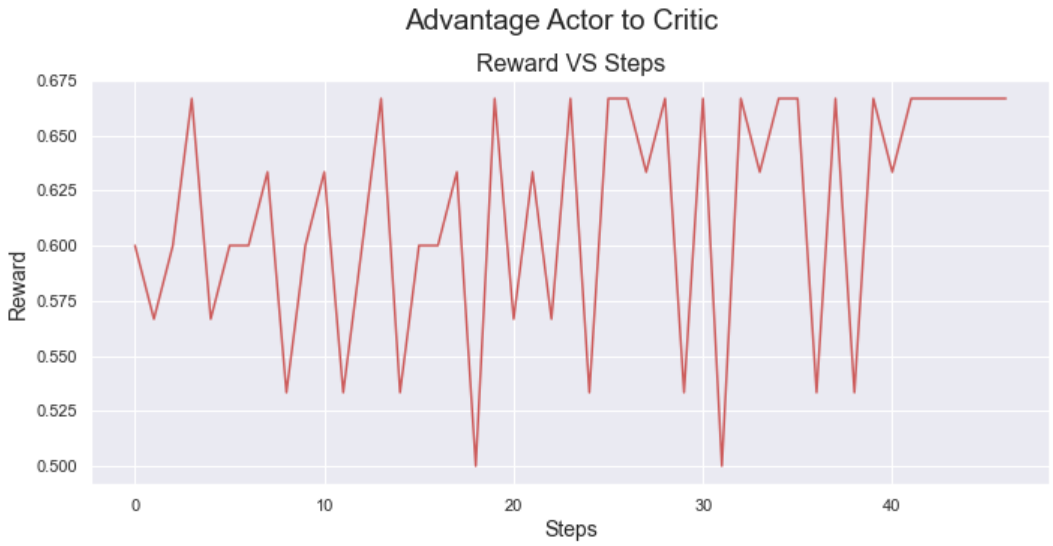


Σχήμα 5.1: *Διαφορετικές τιμές του ρυθμού εκμάθησης για τον A2C (Εκπαίδευση)*

Κατα την αξιολόγηση του πράκτορα, παρατηρείται ότι υπάρχουν αυξομειώσεις στην ανταμοιβή, της τάξης όμως του 0.1 εως 0.2, κάτι το συμβαίνει αρκετά συχνά στον τομέα της βαθιάς ενισχυτικής μάθησης και επιπλέον το γεγονός ότι ξεκινάει η ανταμοιβή του πράκτορα απο το 0.5 και φτάνει περίπου εως το 0.7 ειναι πολύ ενθαρρυντικό. Ο αριθμός των καλά αναπτυγμενων ικανότητων μικρής κλίμακας είναι 25, όπως παρουσιάζεται στον πίνακα 5.1, και το επεισόδιο τερματίζει όταν δεν υπήρχε παραπάνω χρονικός προυπολογισμός. Όπως φαίνεται, η ρεαλιστικότητα του προβλήματος το κάνει ιδιαίτερα ενδιάφερον, αφού ο πράκτορας τερμάτισε το επεισόδιο όταν δεν είχε άλλο χρονικό προυπολογισμό και παρολ' αυτά παρατηρούμε πως η ενισχυτική μάθηση κατέληξε σε έναν αρκετά καλό αριθμό καλά αναπτυγμένων δεξιοτήτων.

Πίνακας 5.1: *Αξιολόγηση του A2C*

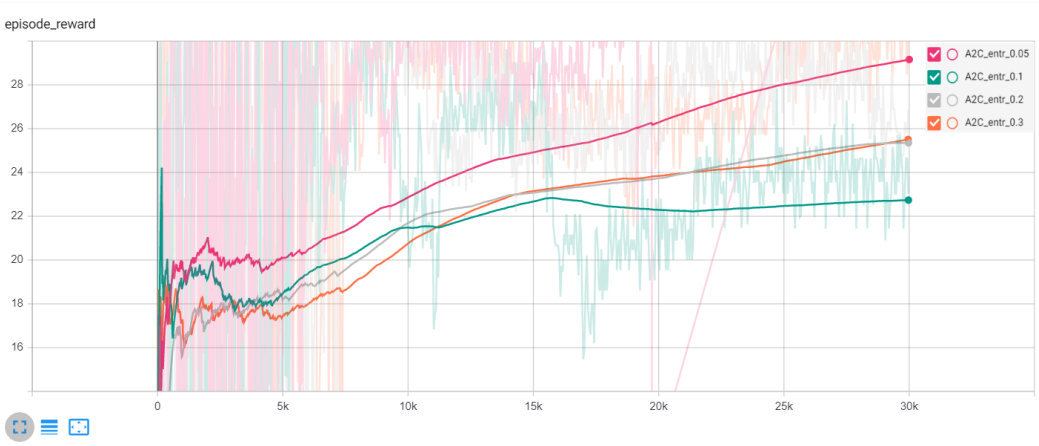
Παράμετρος	**Τιμή**
Ρυθμός εκμάθησης	0.05
Αριθμός βημάτων τερματισμού του επεισοδίου	47
Μέση ανταμοιβή πράκτορα	0.67
Καλά αναπτυγμένες ικανότητες μικρής κλίμακας	25



Σχήμα 5.2: *Ανταμοιβή και Αθροιστική Ανταμοιβή συνάρτησεί των βημάτων για ρυθμο εκμάθησης = 0.05 (Αξιολόγηση)*

Ρύθμιση της εντροπίας

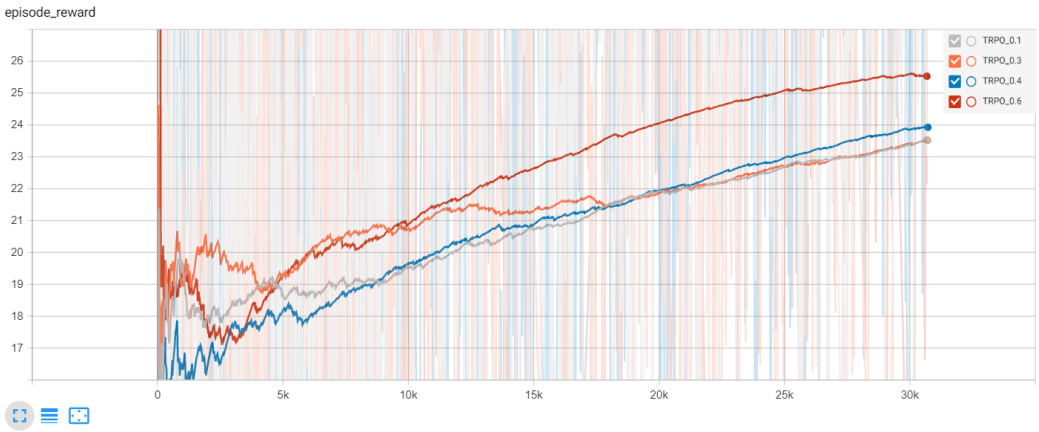
Στα αποτελέσματα που μας δίνει ο Advantage Actor to Critic,παρατηρείται ότι στο στάδιο της αξιολόγησης του πράκτορα, σε κάθε βήμα γίνεται η επιλογή της ίδιας ενέργειας, το οποίο οδηγεί στο συμπέρασμα ότι είτε ο πράκτορας στοχεύει σε μια συγκεκριμένη κατηγορία ικανοτήτων μικρής κλίμακας είτε πιθανώς έχει ¨πέσει¨ σε ένα τοπικό μέγιστο της πολιτικής. Για τον εντοπισμό της αιτίας αυτής της συμπεριφοράς, η επομένη παράμετρος προς βελτιστοποίηση είναι η εντροπία. Για τιμές της παραμέτρου στο διάστημα [0.05,0.1,0.2,0.3], προκύπτει το διάγραμμα 5.3. Για τιμές της εντροπίας ίσες με 0.1,0.2 και 0.3, δηλαδή πολύ έντονη εξερεύνηση του πράκτορα, η ανταμοιβή του είναι χαμηλότερη (εως περίπου το 25%), ενώ για τιμή ίση με 0.05, καταλήγουμε σε ανταμοιβή ίση με 30. Συνεπώς, στο πρόβλημα μας, για τιμές της εντροπίας χαμηλές, λαμβάνουμε μια ικανοποιητική συμπεριφορά και πιο συγκεκριμένη για τιμή ίση με μηδέν (διάγραμμα 5.1) η ανταμοιβή είναι μεγαλύτερη. Συμπεραίνουμε ότι είτε ο συγκεκριμένος αλγόριθμος στοχεύει σε μια συγκεκριμένη κατηγορία ικανοτήτων μικρής κλίμακας της εκπαίδευτικής ομάδας είτε κολλάει σε τοπικά μέγιστα. Με τη χρήση του αλγορίθμου TRPO, ο οποίος μας εξασφαλίζει ότι η πολιτική δεν θα αλλάξει απότομα με αποτέλεσμα να κολλήσει σε ένα τοπικό μέγιστο, ελέγχουμε αν μπορεί να λυθεί το πρόβλημα επιλογής της ίδιας ενέργειας.



Σχήμα 5.3: *Διαφορετικές τιμές της εντροπίας για τον A2C (Εκπαίδευση)*

5.2.2 Trust Policy Optimization

Όπως και προηγουμένως, η πρώτη παραμέτρος προς βελτιστοποίηση του συγκεκριμένου αλγορίθμου είναι η Kullback-Leibler divergence, η οποία μετράει την απόσταση μεταξύ της παλιάς και της νέας πολιτικής. Οι τιμές που δοκιμάζονται για αυτή την παράμετρο είναι 0.1,0.3,0.4 και 0.6. Σύμφωνα με το διάγραμμα 5.4, ο πράκτορας εκπαιδεύεται για τις συγκεκριμένες τιμές της παραμέτρου για 30.000 βήματα, παρουσιάζει σταδιακή βελτίωση στην ανταμοιβή και καταλήγει σε λίγο πιο χαμηλές τιμές της ανταμοιβής.



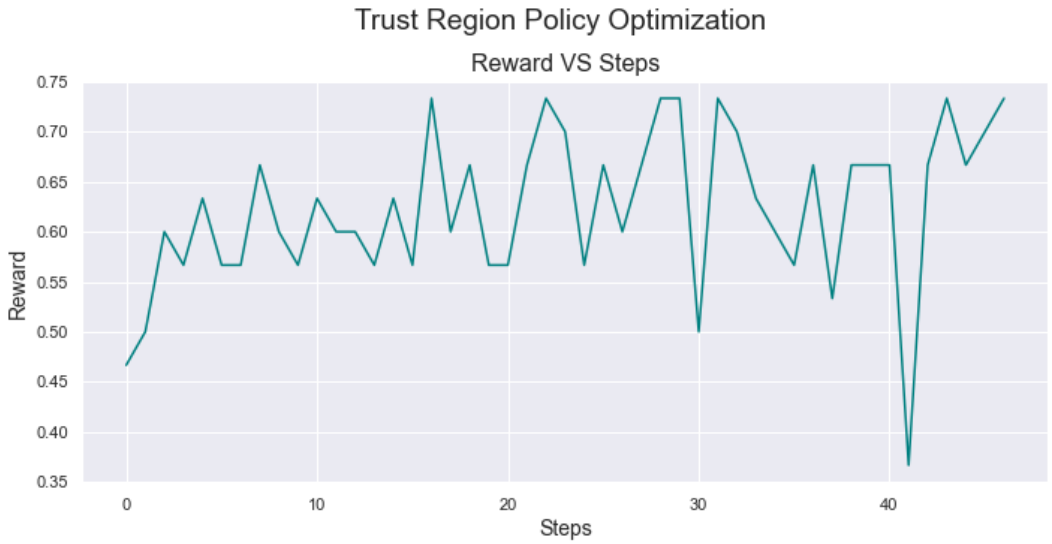
Σχήμα 5.4: *Διαφορετικές τιμές της KL divergence παραμέτρου για τον TRPO (Εκπαίδευση)*

Κατα την αξιολόγηση του πράκτορα, παρατηρείται και πάλι υπάρχουν αυξομειώσεις στην ανταμοιβή, όμως πολύ μικρότερης τάξης πλέον. Επιπλέον, ο πράκτορας επιλέγει διαφορετικές ενέργειες σε κάθε βήμα και ο αριθμός των καλά ανεπτυγμενων ικανότητων μικρής κλίμακας είναι 26, όπως παρουσιάζεται στον πίνακα 5.2. Παρατηρείται δηλαδή μια πολύ καλύτερη συμπεριφορά κατα τη διάρκεια της αξιολόγησης, επιλέγοντας διαφορετικές ενέργειες σε κάθε βήμα και καταλήγοντας σε έναν καλύτερο αριθμό δεξιοτήτων μικρής κλίμακας της εκπαιδευτικής ομάδας. Παρόλο που καταλήγει σε τιμές ανταμοιβής λίγο χαμηλότερες απο αυτές στις οποίες καταλήγει ο Advantage Actor to Critic, φαίνεται ότι στην αξιολόγηση του είναι πιο εμπίστος ως προς τα αποτελέσματα και καταφέρνει να ¨ξεκολλήσει¨ απο ενα

τοπικό μέγιστο της πολιτικής.

Πίνακας 5.2: *Αξιολόγηση του TRPO*

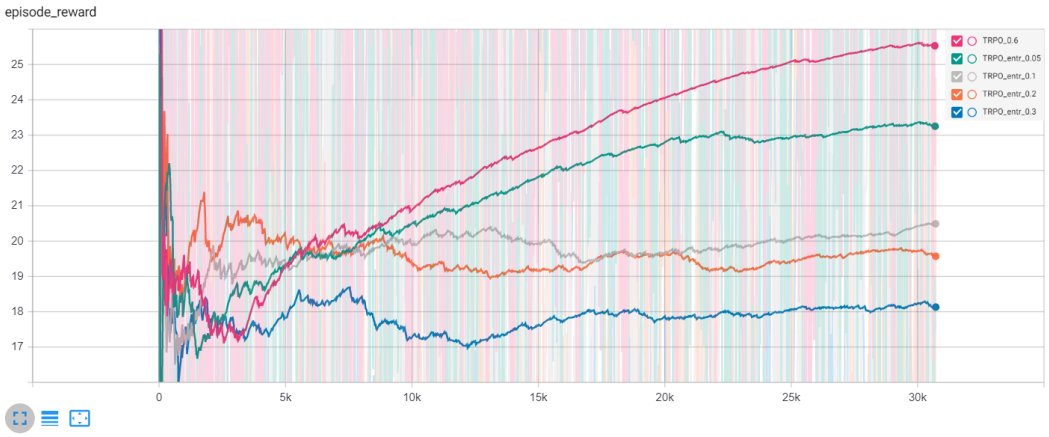
Παράμετρος	**Τιμή**
KL divergence	0.6
Αριθμός βημάτων τερματισμού του επεισοδίου	47
Μέση ανταμοιβή πράκτορα	0.6
Καλά αναπτυγμένες ικανότητες μικρής κλίμακας	26



Σχήμα 5.5: *Ανταμοιβή και Αθροιστική Ανταμοιβή συνάρτησεί των βημάτων για KL divergence = 0.6 (Αξιολόγηση)*

Ρύθμιση της εντροπίας

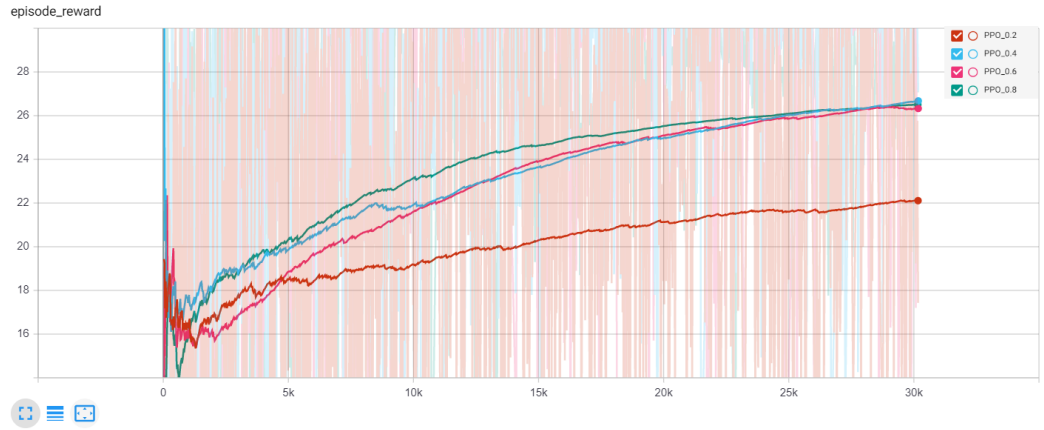
Στη συνέχεια, όπως και πριν, δοκιμάζονται διάφορες τιμές για την παράμετρο της εντροπίας,σε περίπτωση που μπορεί να δώσει καλύτερα αποτελέσματα. Σύμφωνα με το διάγραμμα 5.6, παρατηρείται ότι και πάλι για τιμές της εντροπίας ίσες με 0.1,0.2 και 0.3, δηλαδή πολύ έντονη εξερεύνηση του πράκτορα, η ανταμοιβή του είναι χαμηλότερη ενώ για τιμή ίση με 0.05 είναι λίγο χαμηλότερη απο ό,τι για τιμή ίση με 0 (ρόζ χρώμα). Συνολικά όμως, τα αποτελέσματα είναι ικανοποιητικά χωρίς την εντροπία, με χρήση του συγκεκριμένου αλγορίθμου (βλ. πράσινη και ροζ ανταμοιβή).



Σχήμα 5.6: *Διαφορετικές τιμές της εντροπίας για τον TRPO (Εκπαίδευση)*

5.2.3 Proximal Policy Optimization

Ο τελευταίος αλγόριθμος που χρησιμοποιείται είναι ο Proximal Policy Optimization. Σε αυτόν τον αλγόριθμο η βασική παράμετρος είναι το clip epsilon, η οποία ειναι η υπερπαράμετρος ε που ρυθμίζει το διάστημα στο οποίο ανήκει ο λόγος της παλιάς με τη νέα πολιτική. Εκπαιδεύοντας τον πράκτορα για 30.000 βήματα και δοκιμάζοντας τιμές για την παράμετρο ε στο διάστημα [0.2,0.4,0.6,0.8], προκύπτει το διάγραμμα 5.7,όπου για τιμές μεγαλύτερες απο 0.4 προκύπτουν πολύ καλά αποτελέσματα.



Σχήμα 5.7: *Διαφορετικές τιμές της clip epsilon παραμέτρου για τον PPO (Εκπαίδευση)*

Όμως για τιμή ίση με 0.4, παρατηρείται μια πιο ομαλή αύξηση με τελική τιμή ανταμοιβής ελάχιστα καλύτερη απο την αντίστοιχη που προκύπτει με τις άλλες δύο τιμές.

Κατα την αξιολόγηση του πράκτορα, υπάρχουν αυξομειώσει στην ανταμοιβή και πιο συγκεκριμένα σε σύγκριση με τους δύο άλλους αλγορίθμους είναι οι πιο έντονες αυξομειώσεις. Ο πράκτορας, που έχει εκπαιδευθεί, επιλέγει διαφορετικές ενέργειες αλλα παρατηρείται ότι αυτές οι ενέργειες που επιλέγει μπορεί να διαφέρουν μόνο ως προς μία ικανότητα μικρής κλίμακας που στοχεύουν. Ο αριθμός των καλά αναπτυγμένων ικανότητων μικρής κλίμακας είναι 26,πως παρουσιάζεται στον πίνακα 5.3,και γενικώς έχει μια σχετικά μέτρια απόδοση.

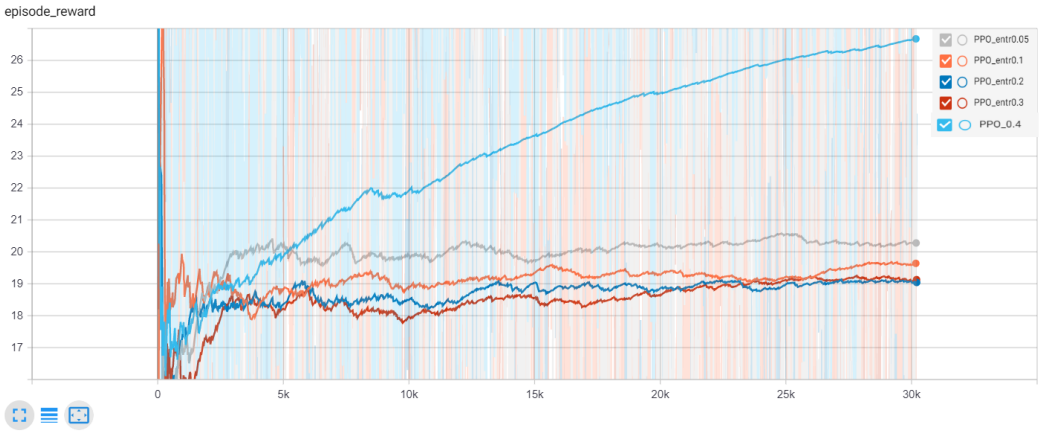
Πίνακας 5.3: *Αξιολόγηση του PPO*

Παράμετρος	**Τιμή**
Clip epsilon	0.4
Αριθμός βημάτων τερματισμού του επεισοδίου	47
Μέση ανταμοιβή πράκτορα	0.6
Καλά αναπτυγμένες ικανότητες μικρής κλίμακας	26

Σχήμα 5.8: *Ανταμοιβή και Αθροιστική Ανταμοιβή συνάρτησεί των βημάτων για clip epsilon = 0.4 (Αξιολόγηση)*

Ρύθμιση της εντροπίας

Τέλος, δοκιμάζοντας διάφορες τιμές για την παράμετρο της εντροπίας,σε περίπτωση που μπορεί να δώσει καλύτερα αποτελέσματα. Σύμφωνα με το διάγραμμα 5.9, παρατηρείται ότι για μη μηδενικές τιμές της εντροπίας υπάρχει μια σημαντική μείωση στην ανταμοιβή. Ενώ για τιμή της εντροπίας ίση με μηδέν (γαλάζια), φαίνεται ότι ο πράκτορας λαμβάνει ενέργειες με πολύ υψηλότερη ανταμοιβή. Αυτή η συμπεριφορά φάνηκε και στο στάδιο της αξιολόγησης, όπου τα αποτελέσματα με μη μηδενική τιμή εντροπίας δεν ήταν τα επιθυμητά.



Σχήμα 5.9: *Διαφορετικές τιμές της εντροπίας για τον PPO (Εκπαίδευση)*

5.2.4 Σύγκριση αλγορίθμων με αύξηση των βημάτων εκπαίδευσης

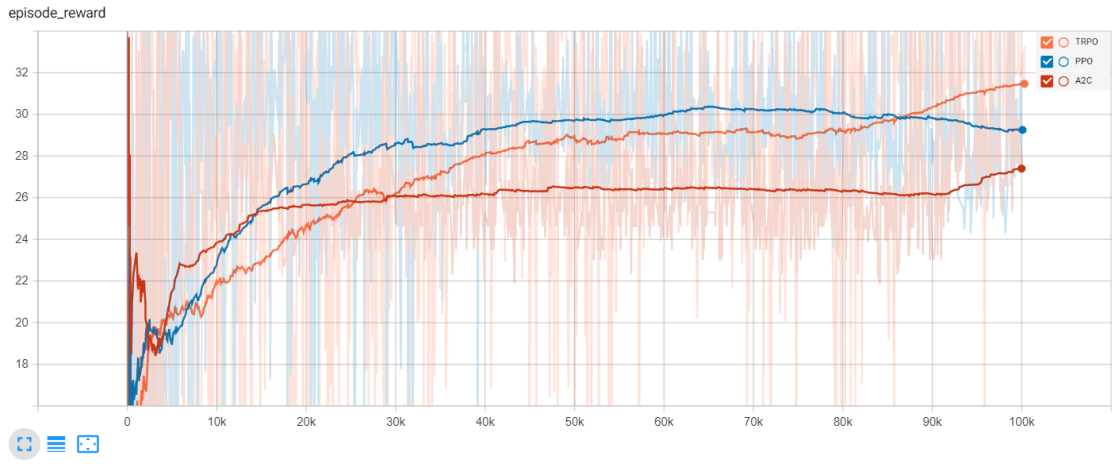
Καταλήγοντας στις παρακάτω παραμέτρους 5.4 για τον κάθε αλγόριθμο, πραγματοποιείται η εκπαίδευση του πράκτορα για μεγαλύτερο αριθμό βημάτων, ίσο με 100.000 βήματα. Επιλέγονται 100.000 βήματα, διότι για κάθε αλγόριθμο παρατηρήθηκε μια ανοδική τάση στην ανταμοιβή και θέλουμε να ελέγξουμε αν αυτή η ανοδική τάση μπορεί να δώσει υψηλότερες τιμές για την ανταμοιβή, χωρίς όμως παράλληλα να απαιτείται ένας τεράστιος αριθμός βημάτων.

Πίνακας 5.4: *Καλύτερες παράμετροι για κάθε αλγόριθμο*

Παράμετρος	**Τιμή**
Learning Rate (A2C)	0.05
KL Divergence (TRPO)	0.6
Clip epsilon (PPO)	0.4

Επιπρόσθετα, για τη σύγκριση των αλγορίθμων χρησιμοποιείται η συνάρτηση EvalCallback της βιβλιοθήκης Stable-Baselines, η οποία αξιολογεί την απόδοση του πράκτροα, χρησιμοποιώντας ένα ξεχωριστό περιβάλλον αξιολόγησης. Μετά απο 500 ¨καλέσματα¨ του ¨callback¨ κατα τη διάρκεια της εκπαίδευσης του πράκτορα, χρησιμοποιείται το περιβάλλον αξιολόγησης το οποιο επιστρέφει επιπλέον τη μέση ανταμοιβή με την τυπική απόκλιση της ανταμοιβής.

Τα αποτελέσματα που προκύπτουν κατα την εκπαίδευση του πράκτορα παρουσιάζονται ακολούθως:



Σχήμα 5.10: *Σύγκριση αλγορίθμων κατα την εκπαίδευση*

Σύμφωνα με το διάγραμμα 5.11, ο αλγόριθμος TRPO παρουσιάζει μια πολύ καλή ανοδική τάση και μετά τα 80.000 βήματα φαίνεται να αυξάνεται κάπως πιο ραγδαία η ανταμοιβή του πράκτορα. Ο αλγόριθμος PPO φαίνεται να παρουσιάζει μια σταθερότητα απο τα 40.000 βήματα εως τα 80.000 και στη συνέχεια να ελλατώνεται ελαφρώς η ανταμοιβή. Τέλος, ο αλγόριθμος A2C δίνει τη χαμηλότερη ανταμοιβή σε σχέση με τους άλλους δύο αλγορίθμους παρουσιάζοντας μια σταθερότητα για ενα μεγάλο εύρος βημάτων.

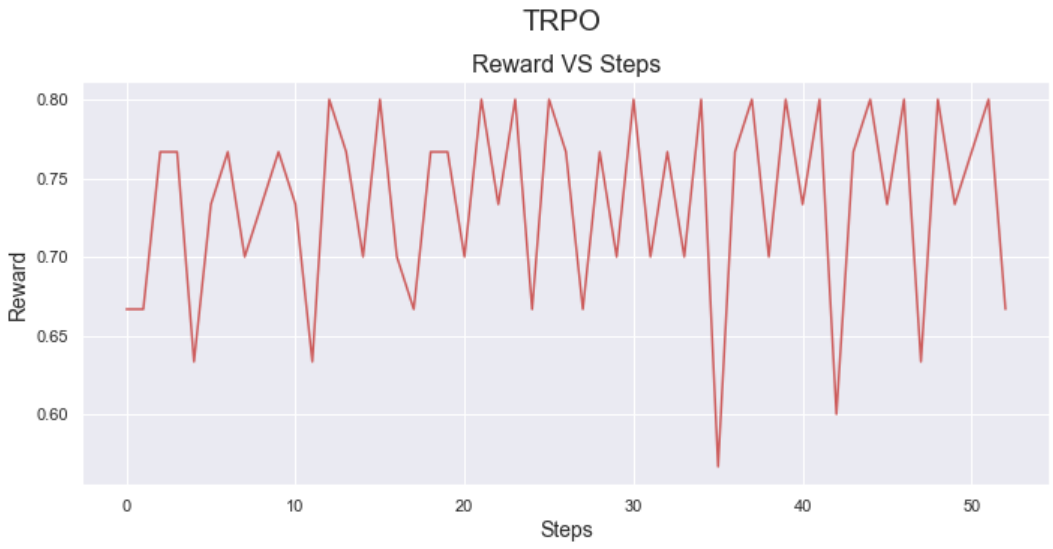
Όσον αφορά την αξιολόγηση του πράκτορα, ενδεικτικά παρουσιάζονται τα αποτελέσματα που μας προσφέρει η συνάρτηση EvalCallback για την εποπτεία της εκπαίδευσης:

```
Eval num_timesteps=500, episode_reward=16.23 +/- 0.75
Episode length: 52.40 +/- 2.58
New best mean reward!
Eval num_timesteps=1000, episode_reward=17.74 +/- 1.10
Episode length: 48.40 +/- 3.01
New best mean reward!
Eval num_timesteps=1500, episode_reward=12.77 +/- 0.58
Episode length: 49.80 +/- 2.48
Eval num_timesteps=2000, episode_reward=20.35 +/- 0.78
Episode length: 47.00 +/- 1.79
New best mean reward!
Eval num_timesteps=2500, episode_reward=25.20 +/- 0.00
Episode length: 33.00 +/- 0.00
New best mean reward!
Eval num_timesteps=3000, episode_reward=29.57 +/- 0.93
Episode length: 48.20 +/- 1.47
```

Σχήμα 5.11: *Αποτελέσματα με τη συνάρτηση EvalCallback*

Μεσω της συνάρτησης EvalCallback, ανά κάποια χρονικά βήματα μπορούμε να αξιολογούμε την εκπαιδεύση του πράκτορα με ενα αντίγραφο του περιβάλλοντος, που ονομάζεται περιβάλλον αξιολόγησης. Μας δίνει τη δυνατότητα να μπορούμε να δούμε τη μέση ανταμοιβή με την τυπική απόκλιση της και το μέσο μήκος επεισοδίου με την τυπική απόκλιση του, αποκτώντας έτσι καλύτερα εποπτεία της εκπαίδευσης του πράκτορα.

Παρατηρώντας τα αποτελέσματα που προκύπτουν κατα την αξιολόγηση, καταλήγουμε σε 23 καλώς αναπτυγμένες ικανότητες μικρής κλίμακας της ομάδας με μέση ανταμοιβή 0.7 έπειτα απο 53 επεισόδια, όπως παρουσιάζονται στο σχήμα 5.12.



Σχήμα 5.12: *Ανταμοιβή και Αθροιστική Ανταμοιβή συνάρτησεί των βημάτων του TRPO με 100κ βήματα*

Επιπλέον με τον συγκεκριμένο αλγόριθμο επιλέγονται διαφορετικές ενέργειες κατα την αξιολόγηση και η ανταμοιβή κυμαίνεται απο 0.65 εως 0.8, δηλαδή φαίνεται τα επιπλέον βήματα να βοήθησαν περαιτέρω τον πράκτορα να εκπαιδευθεί.

Σχετικά με τον αλγόριθμο PPO, που παρουσιάσε τα αμέσως καλύτερα αποτελέσματα κατα την εκπαίδευση,ο πράκτορας καταφέρνει κατα την αξιολόγηση να δώσει μέση ανταμοιβή ίση με 0.6 έπειτα απο 46 βήματα. Συνεπώς, όπως αναμέναμε, η απόδοση του αλγορίθμου TRPO σε σχέση με τους άλλους δύο αλγόριθμους φαίνεται να ειναι ιδιαιτέρως ικανοποιητική, να δίνει πιο σταθερά αποτελέσματα προσφέροντας μια πολιτική που καταφέρνει να προσαρμοστεί όσο καλύτερα γίνεται στο πρόβλημα μας.

Να σημειωθεί ότι στα αποτελέσματα υπάρχει τυχαιότητα στη δημιουργία των καταστάσεων και των ενεργειών, όπως συμβαίνει σε προβλήματα ενισχυτικής μάθησης.

Μέρος III

Επίλογος

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία, υλοποιήθηκε ένα σύστημα σύστασης βασισμένο στην ενισχυτική μάθηση, με στόχο τη σύσταση εκπαιδευτικών δραστηριοτήτων σε ομάδες μαθητών που έχουν ανάγκη για βελτίωση των κοινωνικών και συναισθηματικών τους ικανοτήτων. Το συγκεκριμένο σύστημα υλοποιήθηκε ακολουθεί τη διεπαφή Open AI Gym και εκπαιδεύτηκε με τη βοήθεια αλγορίθμων βαθιάς ενισχυτικής μάθησης της βιβλιοθήκης Stable Baselines.

Στα αποτελέσματα εκπαίδευσης και αξιολόγησης του πράκτορα με τους διαφορετικούς αλγορίθμους προέκυψαν διάφορες παρατηρήσεις και συμπεράσματα. Αρχικά, ο αλγόριθμος Advantage Actor to Critic σε σχέση με τους υπόλοιπους δύο αλγορίθμους, κατά τη διάρκεια εκπαίδευσης του πράκτορα, παρουσίασε υψηλότερες τιμές της ανταμοιβής με το πέρασμα των βημάτων. Να σημειώθει βέβαια ότι αυτή η διαφορά στις τιμές της ανταμοιβής ανάμεσα στους αλγορίθμους δεν ήταν μεγάλου μεγέθους. Όμως, παρόλο που παρατηρήθηκε αυτή η συμπεριφορά για τον συγκεκριμένο αλγόριθμο, κατα τη διάρκεια της αξιολόγησης πρότεινε συνεχώς την ίδια εκπαιδευτική δραστηριότητα με αποτελέσμα να συμπεράνουμε ότι ο αλγόριθμος Advantage A2C ¨κολλούσε¨ σε τοπικά μέγιστα.

Σε αυτό το σημείο, παρουσιάστηκε ο αλγόριθμος TRPO, τον οποίο θα μπορούσαμε να τον περιγράψουμε σαν μια βελτιωμένη προσέγγιση του A2C, ο οποίος μας εξασφαλίζει ότι η πολιτική δεν θα αλλάξει απότομα, με αποτελέσμα ο πράκτορας να μην μπορεί να ¨ρίξει¨ την πολιτική σε τοπικά μέγιστα. Με τα αποτελέσμα που λάβαμε, ο συγκεκριμένος αλγόριθμος παρουσιάσε μια αρκετά καλή εκπαίδευση και παράλληλα στο στάδιο της αξιολόγησης επέλεγε διαφορετικές ενέργειες, με αποτελέσμα να προσπαθεί να στοχεύει και σε διαφορετικές ικανότητες μικρής κλίμακας της εκπαιδευτικής ομάδας που έχουν ανάγκη για βελτίωση. Το πλεονέκτημα αυτού του αλγόρίθμου ήταν ότι μας παρείχε πιο σταθερά αποτελέσμα κατα τη διαδικασία δοκιμής τους στο πρόβλημα.

Τέλος, δοκίμαστηκε και ο αλγόριθμος PPO, ο οποίος απλοποιεί το πρόβλημα μεγιστοποίησης της ανταμοιβής με τον περιορισμό που εισαγει ο TRPO. Ο συγκεκριμένος αλγόριθμος παρουσιάσε κάποιες φορές και εκείνος το πρόβλημα επιλογής μιας συγκεκριμένης εκπαιδευτικής δραστηριότητας κατα τη διαδικασία της αξιολόγησης.

6.2 Μελλοντικές Επεκτάσεις

Κατά τη διάρκεια υλοποίησης της διπλωματικής προέκυψαν σκέψεις, σχετικά με το πώς αυτή μπορεί να βελτιστοποιηθεί και να αυξήσει την αξιοπιστία και τις δυνατότητες του συστήματος.

Ενδιαφέρουσες μελλοντικές επεκτάσεις θα μπορούσαν να ήταν:

- Βελτιστοποίηση περαιτέρω υπερπαραμέτρων των αλγορίθμων ενισχυτικής μάθησης με απώτερο στόχο την αύξηση της ανταμοιβής, όπως για παράδειγμα για τον Advantage Actor to Critic βελτιστοποίηση πολλών υπερπαραμέτρων ταυτοχρόνα έτσι ώστε να αποφευχθούν τοπικά μέγιστα της πολιτικής.

- Υλοποίηση του διαδραστικού συστήματος σύστασης με τη βοήθεια διαφορετικών frameworks, όπως το Pytorch και κατ' επέκταση υλοποίηση των αλγόριθμων ενισχυτικής μάθησης.

- Επέκταση του περιβάλλοντος με την είσοδο δεδομένων που παράγονται απο ένα γράφο, ο οποίος χαρτογραφεί τις ικανότητες μιας εκπαιδευτικής ομάδας τόσο σε ατόμικό όσο και σε ομαδικό επίπεδο . Οι ικανότητες μικρής κλίμακας δηλαδή μπορούν να αντικατασταθούν με μετρικές που παράγονται από ικανότήτες συναισθηματικής νοημοσύνης μιας οντολογίας βασισμένης στη συλλογική συναισθηματική νοημοσύνη. Περαιτέρω επεκτάσεις μπορούν να γίνουν και για την ανταμοιβή.

- Δημιουργία γραφικής διεπαφής που προσωμοιώνει τις εκπαιδευτικές ομάδες με τις ικανότητες τους με σκοπό τη σύσταση ενεργειών σε αυτές. Μπορεί να απευθύνεται σε άτομα χωρίς προγραμματιστική εμπειρία, που θέλουν να προσωμοιώσουν τέτοιες εκπαιδευτικές κατάστασεις.

Βιβλιογραφία

[1] *Categories of methods of Recommender Systems*. https://observablehq.com/@sandraviz/recommender-systems.

[2] *Deep Reinforcement Learning for News Recommendation. Part 1: Architecture*. https://towardsdatascience.com/deep-reinforcement-learning-for-news-recommendation-part-1-architecture-5741b1a6ed56.

[3] *A Taxonomy of RL Algorithms*. https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html.

[4] Sergey Levine Michael I. Jordan John Schulman, Philipp Moritz και Pieter Abbeel. *High-Dimensional Continuous Control Using Generalized Advantage Estimation*. *ICLR 2016*, 2015.

[5] Richard S. Sutton και Andrew G. Barto. *Reinforcement Learning:An Introduction*. The MIT Press, Cambridge, Massachusetts London, England, 2η έκδοση, 2014,2015.

[6] *How does an A3C work?* https://medium.com/@dmonn/how-does-an-a3c-work-4e02266d1a96.

[7] *Policy Gradient Algorithms*. https://lilianweng.github.io/posts/2018-04-08-policy-gradient/.

[8] *Trust Region Policy Optimisation(TRPO) — a policy-based Reinforcement Learning*. https://medium.com/intro-to-artificial-intelligence/trust-region-policy-optimisation-trpo-a-policy-based-reinforcement-learning-fd38ff9e996e.

[9] Prafulla Dhariwal Alec Radford Oleg Klimov John Schulman, Filip Wolski. *Proximal Policy Optimization Algorithms*. 2017.

[10] P´erez Gonz´alez J.C. Garcia Navarro E Bisquerra Alzina, R. *Inteligencia emotional en educaci´on*. Editorial S´ıntesis, 2015.

[11] *Reinforcement Learning Agents*. https://es.mathworks.com/help/reinforcement-learning/ug/create-agents-for-reinforcement-learning.html.

[12] Haifeng Chen Guofei Jiang Hui Zhang Kenji Yoshihira. *Boosting the Performance of Computing Systems throughAdaptive Configuration Tuning*. *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC)*, 2009.

[13] Michalis Feidakis Dimitrios Metafas Symeon Papavassiliou Eleni Fotopoulou, Anastasios Zafeiropoulos. *An Interactive Recommender System Based on Reinforcement Learning for Improving Emotional Competences in Educational Groups*. *In book: Intelligent Tutoring Systems*, σελίδες 248–258, 2020.

[14] Peter Salovey John D. Mayer, David R. Caruso. *The ability model of emotional intelligence:Principles and updates*. *Emotion Review 8(4)*, σελίδες 290–300, 2016.

[15] K. V. Petrides. *Ability and trait emotional intelligence*. *The Wiley-Blackwell handbook of individual differences (pp. 656-678)*, 2013.

[16] D. Goleman. *Emotional intelligence: why it can matter more than IQ*. *Bantam*, 1996.

[17] Folajimi Y. Ojokoh B. Isinkaye, F. *Recommendation systems: Principles, methods and evaluation*. *Egyptian Informatics Journal*, 16(9):261–273, 2015.

[18] *Introduction to recommender systems*. https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada.

[19] Á. Rocha και T. Guarda. *Recommendation Systems in Education:A Systematic Mapping Study*. *Proceedings of the International Conference on InformationTechnology Systems (ICITS 2018), Advances in Intelligent Systems and Computing 721*, 2018.

[20] Prasad P. W. C. Alsadoon A. Maag A. Khanal, S. S. *A systematic review: machine learning based recommendation systems for e-learning*. *Springer Science+Business Media, LLC, part of Springer Nature 2019*, 2019.

[21] Aixin Sun Shuai Zhang, Lina Yao και Yi Tay. *Deep Learning based Recommender System: A Survey and New Perspectives*. *ACM Comput. Surv. 1, 1, Article 1*, σελίδες 1–38, 2019.

[22] Jing Wang Sanmit Narvekar Ritesh Agarwal Rui Wu Heng Tze Cheng Morgane Lustman Vince Gatto Paul Covington Jim McFadden Tushar Chandra Craig Boutilier Eugene Ie, Vihan Jain. *Reinforcement Learning for Slate-based Recommender Systems: A Tractable Decomposition and Practical Methodology*. *Google Research*, 2019.

[23] Ronen I Brafman. Guy Shani, David Heckerman. *An MDP-based Recommender System*. *Journal of Machine Learning Research*, 2002.

[24] *Recommendation Systems using Reinforcement Learning*. RecommendationSystemsusingReinforcementLearning. Ημερομηνία πρόσβασης: 31-12-2019.

[25] Adi Botea Oznur Alkan, Elizabeth M. Daly. *An Evaluation Framework for Interactive Recommender System*. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization - UMAP'19 Adjunct.*, 2019.

[26] *Part 2: Kinds of RL Algorithms*. https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html.

[27] *Reinforcement Learning algorithms — an intuitive overview*. https://smartlabai.medium.com/reinforcement-learning-algorithms-an-intuitive-overview-904e2dff5bbc.

[28] Abay R. Abbey J. Balage S. Brown M. Boyce R. Smith, B. *Propulsionless Planar-Phasing of Multiple Satellites using Deep Reinforcement Learning*. *Advances in Space Research*, 2020.

[29] *Policy Gradients in a Nutshell*. https://towardsdatascience.com/policy-gradients-in-a-nutshell-8b72f9743c5d.

[30] Sergios Karagiannakos. *Deep Reinforcement Learning Course*. AI-Summer, 1η έκδοση, 2018.

[31] Mehdi Mirza Alex Graves Timothy P. Lillicrap Tim Harley David Silver Koray Kavukcuoglu Volodymyr Mnih, Adrià Puigdomènech Badia. *Asynchronous Methods for Deep Reinforcement Learning*. *ICML 2016*, 2016.

[32] Philipp Moritz Michael I. Jordan Pieter Abbeel John Schulman, Sergey Levine. *Trust Region Policy Optimization*. *ICML 2015*, 2015.

[33] *Understanding Proximal Policy Optimization (Schulman et al., 2017)*. https://towardsdatascience.com/understanding-and-implementing-proximal-policy-optimization-schulman-et-al-2017-9523078521ce.

[34] Ronald J και Jing. Peng. *Function Optimization Using Connectionist Reinforcement Learning Algorithms*. *Connection Science*, 1991.

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
κατασταση	state
ανταμοιβή	reward
στρατηγική	policy
πολιτική	policy
περιβάλλον	environment
ενέργεια	action
δραστηριότητα	action
ικανότητα μικρής κλίμακας	micro-competence
μέθοδος κλίσης	gradient descent method