ΕΚΠΑΙΔΕΥΣΗ ΕΥΦΥΗ ΠΡΑΚΤΟΡΑ ΣΤΟ ΠΕΡΙΒΑΛΛΟΝ ΤΟΥ GOOGLE RESEARCH FOOTBALL ΜΕ ΤΗ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΒΑΘΙΑΣ ΕΝΙΣΧΥΤΙΚΗΣ ΜΑΘΗΣΗΣ

Συγγραφέας

Γεώργιος Μουτσόπουλος

Επιβλέπων Καθηγητής

Κωνσταντίνος Μπλέκας

ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA

Σεπτέμβριος 2022

Πίνακας Περιεχομένων

Περίληψη

Τα τελευταία χρόνια έχει παρατηρηθεί αυξημένη πρόοδος στον τομέα της ενισχυτικής μάθησης με τη χρήση εικονικών περιβαλλόντων όπως τα βιντεοπαιχνίδια, όπου καινούργιες μεθοδολογίες και ιδέες μπορούν να δοκιμαστούν με ευκολία και ασφάλεια.

Στα πλαίσια της παρούσας διπλωματικής εργασίας γίνεται ανάλυση των αλγορίθμων βαθιάς ενισχυτικής μάθησης που χρησιμοποιούνται για την εκπαίδευση ευφυών πρακτόρων για την επίλυση διαφόρων προβλημάτων. Ακόμα, γίνεται εφαρμογή τέτοιων αλγορίθμων για την εκπαίδευση ενός ευφυούς πράκτορα στο περιβάλλον 'Google Research Football'.

Αρχικά γίνεται εκμάθηση ενός εικονικού παίκτη ποδοσφαίρου που ξεκινάει από το κέντρο ενός γηπέδου χωρίς αντίπαλο και έχει σκοπό να βάλει goal στο απέναντι τέρμα. Έπειτα γίνεται εκμάθηση ενός άλλου παίκτη ποδοσφαίρου που ξεκινάει από το κέντρο ενός γηπέδου, με στόχο αυτή τη φορά να βάλει goal στο απέναντι τέρμα, ενώ παράλληλα αντιμετωπίζει έναν αντίπαλο παίκτη.

Λέξεις Κλειδιά: Μηχανική μάθηση ,Ενισχυτική μάθηση, Αλγόριθμος , Βαθιά Νευρωνικά Δίκτυα

Abstract

Recent years have seen an increased progress in the field of reinforcement learning with use of virtual environments such as video games, where new methodologies and ideas can be tested easily and safely.

In the context of this thesis, an analysis concerning the field of deep reinforcement learning algorithms, used for training agents, will take place. Afterwards, such algorithms are applied to instruct an intelligent agent in the 'Google Research Football' environment.

Firstly, the objective is to train a virtual football player, who starts from the center of a football field aiming to score a goal without a rival in the opponent's end line. Then, the objective is to train a different virtual football player, who starts from the center of a football field but this time the goal is to score in the opponent's end line while also having the task of beating the opposition.

Keywords: Machine learning, Reinforcement learning, Algorithm, Deep Neural Networks

Κεφάλαιο 1. Εισαγωγή

1.1 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση είναι ένας κλάδος της μηχανικής μάθησης ο οποίος ασχολείται με το να αντιστοιχίζει καταστάσεις(states) σε ενέργειες(actions). Ο μαθητευόμενος δεν διδάσκεται ποιες ενέργειες να ακολουθήσει , αλλά τις ανακαλύπτει κάνοντάς τες αρχικά τυχαία και έπειτα συγκρίνοντας την ανταμοιβή που λαμβάνει στο τέλος της κάθε μιας. Σε πολλές περιπτώσεις μια ενέργεια μπορεί όχι μόνο να επηρεάσει την τρέχουσα κατάσταση, αλλά και τις επόμενες, οδηγώντας τον πράκτορα στο επιθυμητό αποτέλεσμα και τη λύση του συνολικού προβλήματος [1].Η μέθοδος που χρησιμοποιείται είναι της δοκιμής και λάθους(trial and error), η οποία μετά από διαδοχικές επαναλήψεις οδηγεί στην εύρεση του βέλτιστου τρόπου λύσης του εκάστοτε προβλήματος.

Πιο αναλυτικά, στην ενισχυτική μάθηση αρχικά ορίζεται το περιβάλλον εκπαίδευσης του πράκτορα μέσα στο οποίο θα λάβει χώρα η εκπαίδευση. Ύστερα, επινοείται μια μέθοδος επιβράβευσης(reward function) των επιθυμητών συμπεριφορών και τιμωρίας των ανεπιθύμητων κατά τη διάρκεια της εκμάθησης. Αυτή η μέθοδος αποδίδει θετικές τιμές στις προσδοκώμενες ενέργειες και ενθαρρύνει τον πράκτορα να τις επαναλάβει , ενώ αντίστοιχα αποδίδει αρνητικές τιμές στις ανεπιθύμητες ενέργειες κάνοντας τον να μην τις επιλέξει ξανά .Εν τέλει, με αυτό τον τρόπο ο πράκτορας προγραμματίζεται να αναζητά τη μέγιστη συνολική ανταμοιβή, που μακροπρόθεσμα έχει ως αποτέλεσμα την επίλυση του προβλήματος.

1.1.1 Διαφορές Ενισχυτικής Μάθησης με τη Μάθηση με Επίβλεψη

Η μάθηση με επίβλεψη(supervised learning) είναι ένας κλάδος της μηχανικής μάθησης για την εκμάθηση μιας συνάρτησης που αντιστοιχίζει μια είσοδο σε μια έξοδο. Η δημιουργία αυτής βασίζεται σε ένα σύνολο από γνωστά παραδείγματα. Ο στόχος αυτού του είδους μάθησης είναι το σύστημα να λειτουργεί σωστά σε καταστάσεις που δεν υπάρχουν σαν είσοδοι στο σετ εκπαίδευσης [2].Ο συγκεκριμένος τρόπος μάθησης είναι ο πιο διαδεδομένος, ωστόσο δεν είναι επαρκής για να γίνει μάθηση με αλληλεπίδραση. Στα διαδραστικά προβλήματα δεν είναι δυνατό να λαμβάνονται παραδείγματα επιθυμητής συμπεριφοράς ώστε είναι ταυτόχρονα σωστά και αντιπροσωπευτικά για όλες τις καταστάσεις που πρέπει να ενεργήσει ο πράκτορας. Για την επίλυση τέτοιων προβλημάτων ο πράκτορας πρέπει να **μαθαίνει από τη δική του εμπειρία** μέσω της αλληλεπίδρασης του με το περιβάλλον.

1.1.2 Διαφορές Ενισχυτικής Μάθησης με τη Μάθηση χωρίς Επίβλεψη

Η μάθηση χωρίς επίβλεψη(unsupervised learning) είναι ένας τύπος αλγορίθμου που μαθαίνει μοτίβα για δεδομένα τα οποία δεν έχουν κατηγοριοποιηθεί και δεν έχουν ετικέτα(label). Στόχος είναι η ανακάλυψη μοτίβων και η εύρεση δομών που είναι κρυμμένες μέσα σε αυτά τα δεδομένα. Αν και η μάθηση χωρίς επίβλεψη δεν βασίζεται σε χρήση σωστών παραδειγμάτων και δεν απαιτεί σετ εκπαίδευσης όπως και η ενισχυτική μάθηση , οι διαφορές τους με την ενισχυτική μάθηση έγκεινται στο ότι η μια προσπαθεί να εντοπίσει κάποια κρυφή δομή μέσα στα δεδομένα(μάθηση χωρίς επίβλεψη), ενώ η άλλη προσπαθεί να μεγιστοποιήσει ένα σήμα ανταμοιβής(ενισχυτική μάθηση).

1.2 Η μηχανική μάθηση στα Βιντεοπαιχνίδια

Η ενισχυτική μάθηση και τα βιντεοπαιχνίδια έχουν μεγάλη και αμοιβαία επωφελή ιστορία. Από τη μια πλευρά, τα παιχνίδια είναι πλούσια σε απαιτητικές δοκιμασίες στις οποίες μπορούν να δοκιμαστούν αλγόριθμοι ενισχυτικής μάθησης. Από την άλλη , σε πολλά παιχνίδια οι καλύτεροι παίκτες χρησιμοποιούν την ενισχυτική μάθηση για να βελτιωθούν και να ανακαλύψουν καινούργιους τρόπους προσέγγισης του παιχνιδιού.

Βέβαια , χωρίς τροποποιήσεις οι βασικοί αλγόριθμοι σπάνια είναι επαρκείς για να επιτευχθεί η νίκη σε παιχνίδια υψηλής πολυπλοκότητας [3]. Για αυτό το λόγο είναι

απαραίτητη η τροποποίηση αυτών βάσει του περιβάλλοντος που χρησιμοποιούνται και η καθοδήγησή τους μέσω της συνάρτησης ανταμοιβής.

Τέλος, ιδιαίτερη απήχηση είχε ο αλγόριθμος ενισχυτικής μάθησης ο οποίος χρησιμοποιήθηκε για πρώτη φορά το 2013 για να κερδίσει 7 βιντεοπαιχνίδια της σειράς 'Atari' .Ο αλγόριθμος DQN χρησιμοποίησε για πρώτη φορά βαθιά νευρωνικά δίκτυα συνδυάζοντας τα με τη συνάρτηση Q-Value .Η δημοσίευση του αλγορίθμου αποτέλεσε άλλη μια απόδειξη πως τα μοντέλα ενισχυτικής μάθησης έχουν την ικανότητα να κυριαρχούν σε ένα μεγάλο σύνολο προκλήσεων στο χώρο των βιντεοπαιχνιδιών.

1.3 Αντικείμενο της Διπλωματικής Εργασίας

Στόχος της διπλωματικής εργασίας είναι η μελέτη του περιβάλλοντος 'Google Research Football' και η εκπαίδευση δύο διαφορετικών πρακτόρων σε παρεμφερή περιβάλλοντα με τη χρήση του αλγορίθμου Deep Q Network(DQN).

Στην πρώτη περίπτωση ο ευφυής πράκτορας είναι ένας παίκτης ποδοσφαίρου ο οποίος ξεκινάει από το κέντρο του γηπέδου και προσπαθεί να βάλει goal στο απέναντι άδειο τέρμα, χωρίς να έρχεται αντιμέτωπος με κάποιο εμπόδιο και χωρίς να βγει η μπάλα εκτός ορίων.

Στην δεύτερη περίπτωση ο παίκτης ξεκινάει πάλι από το κέντρο του γηπέδου και αυτή τη φορά πρέπει να περάσει τον αντίπαλο παίκτη και να σκοράρει χωρίς να βγει η μπάλα εκτός ορίων ή να πάει στο άλλο μισό του γηπέδου .

Κεφάλαιο 2. Περιβάλλον Google Research Football

2.1 Διαφορετικές Αναπαραστάσεις του Χώρου Καταστάσεων

Η αναπαράσταση του χώρου καταστάσεων(state space) και του χώρου ενεργειών (action space) μπορεί να γίνει με 3 τρόπους στο περιβάλλον Google Research Football:

1ος τρόπος : Αναπαράσταση με pixels

Η αναπαράσταση αυτή αποτελείται από μια εικόνα 1280×720 RGB που αντιστοιχεί στην οθόνη που γίνεται render.Αυτή η εικόνα περιλαμβάνει το μέρος του γηπέδου και τους παίκτες που βρίσκονται στο σημείο της μπάλας εκείνη τη στιγμή .Ταυτόχρονα στο πάνω μέρος της απεικονίζεται ένας πίνακας με το σκορ, ενώ στο κάτω μέρος υπάρχει ένας μικρός πίνακας με τις θέσεις όλων των παικτών που υπάρχουν στο γήπεδο εκείνη τη στιγμή(εικόνα 1).

Εικόνα 1.Παραδείγματα απεικόνισης του γραφικού περιβάλλοντος με την επιλογή της αναπαράστασης με pixels.

2ος τρόπος : Αναπαράσταση με Super Mini Map(SMM)

Η αναπαράσταση αυτή είναι βασικά μια στοίβα δυαδικών πινάκων που ορίζουν τον μικρό πίνακα-χάρτη που αποδίδεται στο κάτω μέρος της οθόνης. Η αναπαράσταση SMM αποτελείται από τέσσερις πίνακες 96 × 72 που κωδικοποιούν πληροφορίες σχετικά με τη γηπεδούχο ομάδα, τη φιλοξενούμενη ομάδα, τη μπάλα και τον ενεργό παίκτη αντίστοιχα. Η κωδικοποίηση είναι δυαδική, αντιπροσωπεύοντας εάν υπάρχει παίκτης, μπάλα ή ενεργός παίκτης στις αντίστοιχες συντεταγμένες ή όχι (εικόνα 2).

Εικόνα 2.Παράδειγμα απεικόνισης του γραφικού περιβάλλοντος με την επιλογή της αναπαράστασης Super Mini Map.

3ος τρόπος : Αναπαράσταση με Πίνακα από μεταβλητές

Ο πίνακας της αναπαράστασης αποτελείται από 115 μεταβλητές τύπου float και έχει την εξής δομή:

- Οι πρώτες 22 μεταβλητές είναι οι συντεταγμένες των θέσεων(στους άξονες x,y) των παικτών της αριστερής ομάδας.
- Οι επόμενες 22 μεταβλητές είναι οι κατευθύνσεις (στους άξονες x,y) των παικτών της αριστερής ομάδας.
- Οι επόμενες 22 μεταβλητές είναι αντίστοιχα συντεταγμένες των θέσεων(στους άξονες x,y) των παικτών της δεξιάς ομάδας.
- Οι επόμενες 22 μεταβλητές είναι οι κατευθύνσεις (στους άξονες x,y) των παικτών της δεξιάς ομάδας.
- Οι επόμενες 3 μεταβλητές (στους άξονες x,y,z) είναι οι συντεταγμένες της θέσης της μπάλας.
- Οι επόμενες 3 μεταβλητές είναι η κατεύθυνση της μπάλας (στους άξονες x,y,z).
- Οι επόμενες 3 μεταβλητές χρησιμοποιούνται για την κωδικοποίηση της κατοχής της μπάλας σε σχέση με την ομάδα (κανένας , αριστερή ομάδα , δεξιά ομάδα).
- Οι επόμενες 11 μεταβλητές χρησιμοποιούνται για την κωδικοποίηση της κατοχής της μπάλας σε σχέση με τον παίκτη(οι αριθμοί 1 ως 11 αναφέρονται στους παίκτες της εκάστοτε ομάδας).

- Οι τελευταίες 7 μεταβλητές χρησιμοποιούνται για την κωδικοποίηση της λειτουργίας του παιχνιδιού. Οι λειτουργίες που υπάρχουν είναι οι εξής: κανονικό παιχνίδι , έναρξη αγώνα , σκοράρισμα, φάουλ , κόρνερ, πέναλτι, οφσάιντ.

2.2 Χώρος Ενεργειών

Ο χώρος ενεργειών αποτελείται από 19 δράσεις και χωρίζεται σε 4 κατηγορίες:

1. **Ενέργειες Αδράνειας:**
    - action_idle , η ενέργεια δεν έχει κάποια λειτουργία.
2. **Ενέργειες Κίνησης του παίκτη:**
    - action_left, κίνηση προς τα αριστερά.
    - action_top_left, κίνηση προς τα πάνω – αριστερά.
    - action_top , κίνηση προς τα πάνω.
    - action_top_right, κίνηση προς τα πάνω – δεξιά.
    - action_right, κίνηση προς τα δεξιά.
    - action_bottom_right, κίνηση προς τα κάτω δεξιά.
    - action_bottom, κίνηση προς τα κάτω .
    - action_bottom_left, κίνηση προς τα κάτω αριστερά.
3. **Ενέργειες Σουτ/Πάσα:**
    - action_long_pass, εκτέλεση μακρινής πάσας σε παίκτη της ίδιας ομάδας. Ο παίκτης στον οποίο θα πάει η μπάλα καθορίζεται αυτόματα με βάση την κατεύθυνση του παίκτη που κάνει την πάσα.
    - action_high_pass, εκτέλεση ψηλής πάσας (παρόμοια με την ενέργεια action_long_pass).
    - action_short_pass, εκτέλεση χαμηλής πάσας (παρόμοια με την ενέργεια action_long_pass).
    - action_shot, εκτέλεση σουτ, πάντα στην κατεύθυνση του αντίπαλου τέρματος.
4. **Άλλες Ενέργειες:**
    - action_sprint, έναρξη σπριντ, ο παίκτης τρέχει πιο γρήγορα αλλά έχει χειρότερο χειρισμό της μπάλας.
    - action_release_direction, επαναφορά της κατεύθυνσης του παίκτη
    - action_release_sprint, σταμάτημα του σπριντ.

- action_sliding, εκτέλεση τάκλιν, ισχύει όταν ο παίκτης δεν έχει την μπάλα.
- action_dribble , έναρξη ντρίμπλας ,ισχύει όταν παίκτης έχει τη μπάλα, κινείται πιο αργά, αλλά είναι πιο δύσκολο να του πάρεις την μπάλα.
- action_release_dribble, σταμάτημα ντρίμπλας.

2.3 Κανόνες Παιχνιδιού Google Research Football

Κανόνες παιχνιδιού για παιχνίδι 11 εναντίον 11 παικτών :

- Το παιχνίδι αποτελείται από δύο ημίχρονα, 45 λεπτών (1500 βήματα) το καθένα.
- Η έναρξη στην αρχή κάθε ημιχρόνου γίνεται από διαφορετική ομάδα, αλλά δεν υπάρχει ανταλλαγή πλευρών (το παιχνίδι είναι πλήρως συμμετρικό).
- Ο πράκτορας ελέγχει έναν μόνο παίκτη στην ομάδα. Ο ελεγχόμενος παίκτης είναι πάντα αυτός που έχει την μπάλα ή αυτός που είναι πιο κοντά στην μπάλα όταν αμύνεται.
- Κερδίζει η ομάδα που σκόραρε τα περισσότερα goal , αν και οι δύο ομάδες έχουν σκοράρει τον ίδιο αριθμό goal ,τότε υπάρχει ισοπαλία.
- Αν κάποιος παίκτης πάρει κόκκινη κάρτα αποβάλλεται. Αν η ομάδα φτάσει τους 7 παίκτες , δεν θα υπάρξει αποβολή παίκτη
- Δεν υπάρχουν αναπληρωματικοί παίκτες.
- Δεν υπάρχει επιπλέον χρόνος μετά τη λήξη των 90 λεπτών(3000 βημάτων)

Κανόνες παιχνιδιού για παιχνίδι παίκτη χωρίς αντίπαλο :

- Το παιχνίδι αποτελείται από 12 λεπτά(400 βήματα).
- Ο παίκτης ξεκινάει από το κέντρο του γηπέδου με την μπάλα και είναι στραμμένος προς την αντίπαλη εστία .
- Το παιχνίδι τελειώνει όταν βγει η μπάλα εκτός των ορίων του γηπέδου, πάει στο πίσω μέρος του γηπέδου, μπει goal ή τελειώσει ο χρόνος.

Κανόνες παιχνιδιού για παιχνίδι ένας εναντίον ενός:

- Το παιχνίδι αποτελείται από 12 λεπτά(400 βήματα).
- Ο παίκτης ξεκινάει από το κέντρο του γηπέδου με την μπάλα και είναι στραμμένος προς την αντίπαλη εστία.

- Ο αντίπαλος παίκτης ξεκινάει από την εστία του(θέση (-1.0,0.0)) και κινείται προς την μπάλα.
- Το παιχνίδι σταματάει όταν η μπάλα βγει εκτός ορίων του γηπέδου, μπει goal σε οποιαδήποτε από τις δύο εστίες ή τελειώσει ο χρόνος .

2.4 Τροποποίηση Χώρου Καταστάσεων και Ενεργειών για την Εργασία

Για την απλοποίηση του προβλήματος και την εξοικονόμηση χρόνου της εκπαίδευσης του πράκτορα **περιορίστηκε ο Χώρος Καταστάσεων και ο Χώρος Ενεργειών** που αναφέρεται στις ενότητες 2.1 και 2.2 και προσαρμόστηκαν στα δεδομένα του εκάστοτε προβλήματος.

2.4.1 Χώροι Καταστάσεων και Ενεργειών για το παιχνίδι του παίκτη χωρίς αντίπαλο

Ο **χώρος καταστάσεων** είναι ένας πίνακας (μεγέθους 8 αριθμών τύπου float) με τις εξής μεταβλητές:

- ✓ 2 μεταβλητές για τη θέση του παίκτη
- ✓ 2 μεταβλητές για τον προσανατολισμό του παίκτη
- ✓ 3 μεταβλητές για τη θέση της μπάλας
- ✓ 3 μεταβλητές για τον προσανατολισμό της μπάλας

Ο **χώρος ενεργειών** είναι ένας πίνακας που αποτελείται από 4 ενέργειες:

- ✓ action_right, κίνηση προς τα δεξιά
- ✓ action_top_right, κίνηση προς τα πάνω – δεξιά
- ✓ action_bottom_right, κίνηση προς τα κάτω δεξιά
- ✓ action_shot, εκτέλεση σουτ στην κατεύθυνση της απέναντι εστίας

Οι **τροποποιήσεις** που έγιναν στο **περιβάλλον Google Research Football** ήταν:

1. Ο παίκτης θα ξεκινάει κάνοντας ένα βήμα προς την μπάλα
2. Όσο ο παίκτης βρίσκεται εκτός της μεγάλης περιοχής (box) να μην μπορεί να κάνει σουτ.
3. Όταν ο παίκτης εκτελέσει σουτ να μην κάνει καμία άλλη κίνηση μέχρι να σταματήσει το παιχνίδι(δηλαδή μέχρι να βγει η μπάλα εκτός ορίων ή να μπει goal).

2.4.2 Χώροι Καταστάσεων και Ενεργειών για το παιχνίδι ένας εναντίον ενός

Ο **χώρος καταστάσεων** είναι ένας πίνακας (μεγέθους 17 αριθμών τύπου float)με τις εξής μεταβλητές:

- ✓ 2 μεταβλητές για τη θέση του παίκτη
- ✓ 2 μεταβλητές για τον προσανατολισμό του παίκτη
- ✓ 2 μεταβλητές για τη θέση του αντιπάλου
- ✓ 2 μεταβλητές για τον προσανατολισμό του αντιπάλου
- ✓ 3 μεταβλητές για τη θέση της μπάλας
- ✓ 3 μεταβλητές για τον προσανατολισμό της μπάλας
- ✓ 3 μεταβλητές για την κωδικοποίηση της κατοχής της μπάλας σε σχέση με την ομάδα

Ο **χώρος ενεργειών** είναι ένας πίνακας που αποτελείται από 9 ενέργειες:

- ✓ action_right, κίνηση προς τα δεξιά
- ✓ action_top_right, κίνηση προς τα πάνω – δεξιά
- ✓ action_bottom_right, κίνηση προς τα κάτω δεξιά
- ✓ action_release_direction, επαναφορά της κατεύθυνσης του παίκτη
- ✓ action_shot, εκτέλεση σουτ στην κατεύθυνση της απέναντι εστίας
- ✓ action_sprint, έναρξη σπριντ
- ✓ action_release_sprint, σταμάτημα του σπριντ
- ✓ action_dribble , έναρξη ντρίμπλας
- ✓ action_release_dribble, σταμάτημα ντρίμπλας

Οι **τροποποιήσεις** που έγιναν στο **περιβάλλον Google Research Football** ήταν:

1. Ο παίκτης θα ξεκινάει κάνοντας ένα βήμα προς την μπάλα
2. Όσο ο παίκτης βρίσκεται εκτός της μεγάλης περιοχής (box) να μην μπορεί να κάνει σουτ.
3. Αν ο παίκτης περάσει κατά πολύ τον αντίπαλο , τότε να μην μπορεί να ξεκινήσει ή να σταματήσει να κάνει ντρίμπλα.
4. Αν γίνει φάουλ από τον αντίπαλο στον παίκτη, το παιχνίδι τερματίζει
5. Όταν ο παίκτης εκτελέσει σουτ να μην κάνει καμία άλλη κίνηση μέχρι να σταματήσει το παιχνίδι(δηλαδή μέχρι να βγει η μπάλα εκτός ορίων ή να μπει goal).
6. Όταν ο αντίπαλος παίκτης κρατήσει την μπάλα για πάνω από 3 δευτερόλεπτα το παιχνίδι να σταματάει.

Κεφάλαιο 3. Ο αλγόριθμος Q Learning

Ο Q learning είναι ένας off-policy αλγόριθμος ενισχυτικής μάθησης που επιδιώκει να βρει την καλύτερη δυνατή ενέργεια, δεδομένης της τρέχουσας κατάστασης που του δίνεται.

Θεωρείται off policy, επειδή η συνάρτηση Q-learning μαθαίνει από εμπειρίες ανεξαρτήτως της πολιτικής που συλλέγονται, όπως για παράδειγμα με την εφαρμογή τυχαίων ενεργειών. Ο αλγόριθμος Q-learning ψάχνει να βρει μια πολιτική που μεγιστοποιεί τη συνολική ανταμοιβή σε κάθε κατάσταση. Το «q» στον τίτλο του αλγορίθμου «Q-learning» σημαίνει ποιότητα(quality). Η ποιότητα σε αυτήν την περίπτωση αντιπροσωπεύει το πόσο χρήσιμη είναι μια δεδομένη ενέργεια για την απόκτηση κάποιας άμεσης ή μελλοντικής ανταμοιβής [4].

3.1 Μαρκοβιανή διαδικασία απόφασης

Ο κύριος στόχος του αλγορίθμου Q-learning είναι να βρει λύσεις για το πρόβλημα της Μαρκοβιανής Διαδικασίας Απόφασης (Markov Decision Process ή αλλιώς MDP).Μια διαδικασία απόφασης Markov ασχολείται με τον στοχαστικό έλεγχο σε διακριτό χρόνο. Δηλαδή, τον **σχεδιασμό μιας χρονικής διαδρομής** που εκτελεί μια επιθυμητή εργασία **σε διακριτό χρόνο**, με το ελάχιστο δυνατό κόστος και χωρίς την παρουσία θορύβου [5].

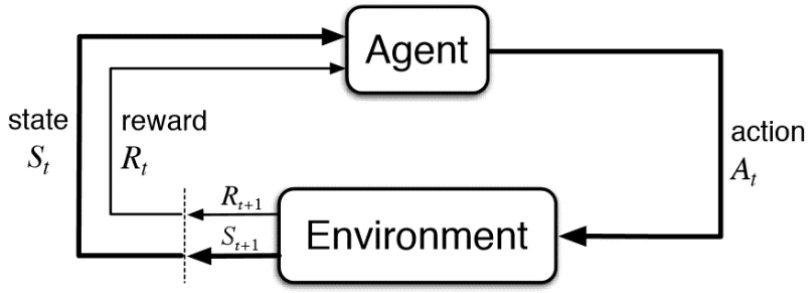
Οι Μαρκοβιανές διαδικασίες απόφασης είναι ιδιαίτερα χρήσιμες για τη μελέτη προβλημάτων βελτιστοποίησης που μπορούν να επιλυθούν μέσω δυναμικού προγραμματισμού.

Στην περίπτωση της ενισχυτικής μάθησης, αυτός που μαθαίνει και παίρνει τις αποφάσεις ονομάζεται **πράκτορας**. Όλο το υπόλοιπο σύστημα με το οποίο ο πράκτορας αλληλοεπιδρά, ονομάζεται **περιβάλλον**. Πράκτορας και περιβάλλον

βρίσκονται σε μια συνεχή αλληλεπίδραση κατά την οποία ο πράκτορας κάνει ενέργειες και το περιβάλλον ανταποκρίνεται σε αυτές.

Κατά τη διάρκεια της εκπαίδευσης, το περιβάλλον επιστρέφει μια τιμή στον πράκτορα η οποία αξιολογεί το πόσο χρήσιμη ήταν μια ενέργεια που εκτέλεσε και ονομάζεται **ανταμοιβή**. Ο πράκτορας λοιπόν διαλέγει τις πράξεις που θα εκτελέσει με βάση την τιμή της ανταμοιβής που θα του επιστρέψει το περιβάλλον[6].

Με αυτό τον τρόπο , μια μαθησιακή συμπεριφορά μπορεί να αναλυθεί με 3 σήματα που εισέρχονται και εξέρχονται μεταξύ του πράκτορα και του περιβάλλοντός του: Ένα σήμα που αντιπροσωπεύει τη βάση/κατάσταση μέσα στην οποία γίνονται οι επιλογές (**state**), ένα σήμα για την αναπαράσταση των επιλογών που έγιναν από τον πράκτορα(**actions**) και ένα σήμα που αναπαριστά την τιμή της ανταμοιβής του πράκτορα(**reward**)(εικόνα 3).



Εικόνα 3. Η Αλληλεπίδραση πράκτορα και περιβάλλοντος σε μια Μαρκοβιανή Διαδικασία Απόφασης.

3.2 Η συνάρτηση Bellman

Η συνάρτηση Bellman εμφανίζεται παντού στη βιβλιογραφία, αποτελώντας ένα από τα κύρια στοιχεία για τη δόμηση των αλγορίθμων ενισχυτικής μάθησης. Ο λόγος για τον οποίο συμβαίνει αυτό είναι επειδή απλοποιεί τον υπολογισμό της **συνάρτησης αξίας** (value function). Έτσι λοιπόν, αντί να αθροίζονται πολλαπλά χρονικά βήματα για την επίλυση ενός προβλήματος, αυτό αναλύεται σε μικρότερα και απλούστερα υπό-προβλήματα, ώστε να βρεθούν οι βέλτιστες λύσεις για το καθένα από αυτά.

Υπάρχουν πολλές διαφορετικές παραλλαγές της συναρτήσεις Bellman. Για την εκπόνηση της παρούσας εργασίας θα χρησιμοποιηθεί η ακόλουθη εξίσωση που φαίνεται στην Εικόνα 4.

$$newQ(s,a) = Q(s,a) + \alpha[R(s,a) + \gamma[maxQ'(s',a') - Q(s,a)]]$$

Εικόνα 4.Συνάρτηση Bellman για την αξιολόγηση μιας ενέργειας

Το **newQ(s,a)** είναι η νέα τιμή που υπολογίζεται για την εισαγωγή του τρέχοντος ζεύγους κατάστασης-ενέργειας (action-value) στον πίνακα Q . Το **Q(s,a)** είναι η τιμή της τρέχουσας κατάστασης, το **a**(alpha) είναι ο ρυθμός εκμάθησης , το **R(s,a)**είναι η τιμή της ανταμοιβής για την συγκεκριμένη ενέργεια στη δοσμένη κατάσταση , το **γ**(gamma) η μεταβλητή μείωσης (discount factor) και το **maxQ'**(s',a') είναι η μέγιστη αναμενόμενη μελλοντική τιμή που είναι πιθανό να δοθεί στην επόμενη κατάσταση.

Ο πίνακας Q μέσα στον οποίο γίνονται οι εισαγωγές ζεύγους κατάστασης – ενέργειας είναι ένας πίνακας δύο διαστάσεων με αριθμό γραμμών ίσο με τον αριθμό των καταστάσεων και αριθμό στηλών ίσο με αυτόν των ενεργειών .Ο πίνακας αρχικοποιείται με όλες τις τιμές του να είναι ίσες με μηδέν και στη συνέχεια ενημερώνεται με τις τιμές Q που προκύπτουν από τη συνάρτηση Bellman (εικόνα 5).

Initialized

Q-Table		Actions					
		South (0)	North (1)	East (2)	West (3)	Pickup (4)	Dropoff (5)
States	0	0	0	0	0	0	0
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	327	0	0	0	0	0	0
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	499	0	0	0	0	0	0

Training

Q-Table		Actions					
		South (0)	North (1)	East (2)	West (3)	Pickup (4)	Dropoff (5)
States	0	0	0	0	0	0	0
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	328	-2.30108105	-1.97092096	-2.30357004	-2.20591839	-10.3607344	-8.5583017
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	499	9.96984239	4.02706992	12.96022777	29	3.32877873	3.38230603

Εικόνα 5.Παράδειγμα πίνακα Q κατά την αρχικοποίηση του και ύστερα μετά την εισαγωγή των τιμών Q.

3.3 Ανάλυση υπέρ-παραμέτρων της συνάρτησης Bellman

Οι πιο σημαντικές παράμετροι για την εκπαίδευση ενός πράκτορα είναι οι εξής:

1. **Alpha**, ρυθμός εκμάθησης (learning rate, lr)
2. **Gamma**, ρυθμός μείωσης (discount rate)
3. **Epsilon** , ρυθμός εξερεύνησης (exploration rate)

3.3.1 Υπέρ-παράμετρος Alpha

Ο **ρυθμός εκμάθησης** Alpha κυμαίνεται από 0 ως 1 και καθορίζει το κατά πόσο οι ανταμοιβές που λαμβάνει ο πράκτορας για μια δεδομένη κατάσταση θα επηρεάσουν τις επόμενες ενέργειες του. Για παράδειγμα αν ο ρυθμός εκμάθησης είναι ίσος με το 0, τότε ο πράκτορας δεν θα μάθει τίποτα από την αλληλεπίδραση του με το περιβάλλον και θα συνεχίσει να παίρνει αποφάσεις τυχαία. Στην περίπτωση που ο ρυθμός εκμάθησης είναι ίσος με το 1 , ο πράκτορας θα καταλήξει να μάθει πολιτικές οι οποίες αφορούν ένα συγκεκριμένο ντετερμινιστικό περιβάλλον, δηλαδή ενέργειες που θα κάνει θα βασίζονται εξ ολοκλήρου στις αρχικές συνθήκες-καταστάσεις και δεν θα υπάρχει καμία τυχαιότητα. Στη δεύτερη περίπτωση το αποτέλεσμα θα είναι να γίνονται για τις ίδιες καταστάσεις ακριβώς οι ίδιες ενέργειες.

3.3.2 Υπέρ-παράμετρος Gamma

Ο ρυθμός μείωσης Gamma κυμαίνεται και αυτός από 0 ως 1 και σχετίζεται με το κατά πόσο μια άμεση ανταμοιβή επηρεάζει την συμπεριφορά του πράκτορα σε σύγκριση με μια μελλοντική. Ο πράκτορας ιδανικά θα πρέπει να είναι σε θέση να επιλέγει μια ενέργεια όχι μόνο βάσει της ανταμοιβής που λαμβάνει από την κατάσταση στην οποία συμβαίνει η ενέργεια, αλλά και βάσει της ανταμοιβής που περιμένει να λάβει και από τις μελλοντικές καταστάσεις, δεδομένων των ενεργειών που θα επιλέξει. Για παράδειγμα, αν επιλεχθεί μια τιμή του Gamma κοντά στο μηδέν, τότε ο πράκτορας δεν θα ενδιαφέρεται για τις μελλοντικές ανταμοιβές και θα λάβει υπόψιν μόνο τις τρέχουσες. Αντίθετα , αν επιλεχθεί μια τιμή κοντά στο 1 , ο πράκτοράς θα θεωρήσει τις μελλοντικές ανταμοιβές ίσης αξίας με τις τρέχουσες(εικόνα 6).

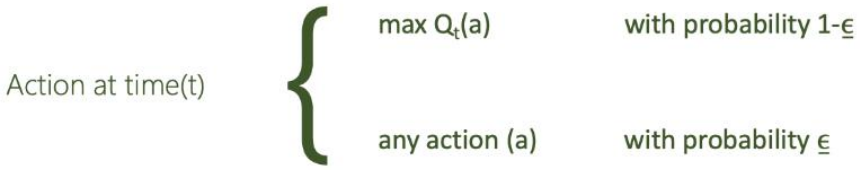
$$\gamma[maxQ'(s',a') - Q(s,a)]$$

Εικόνα 6. Τιμές που επηρεάζει το Gamma στη συνάρτηση Bellman

3.3.3 Υπέρ-παράμετρος Epsilon

Ο **ρυθμός εξερεύνησης** epsilon επίσης κυμαίνεται από 0 ως 1. Καθώς ο πράκτορας εξερευνά το περιβάλλον, μαθαίνει ότι μερικές ενέργειες είναι καλύτερες σε σύγκριση με κάποιες άλλες, για ορισμένες καταστάσεις . Σε πολλές περιπτώσεις , υπάρχει μεγάλη πιθανότητα ο πράκτορας να κολλήσει σε τοπικό μέγιστο, κάνοντας τις ίδιες ενέργειες που νομίζει ότι έχουν τη μέγιστη αξία, ξανά και ξανά, ενώ την ίδια στιγμή μπορεί να υπάρχουν «καλύτερες» ενέργειες που δεν τις έχει ανακαλύψει ακόμα.

Με την επιλογή της τιμής **epsilon**, η ενέργεια που θα επιλέξει ο πράκτορας έχει πιθανότητα ίση με epsilon ότι θα είναι τυχαία, κάτι που οδηγεί στην περεταίρω εξερεύνηση του περιβάλλοντος (exploration). Ταυτόχρονα, υπάρχει πιθανότητα ίση με 1-epsilon ότι η επόμενη ενέργεια που θα πραγματοποιηθεί θα είναι αυτή με την υψηλότερη τιμή Q για την δεδομένη κατάσταση (exploitation).

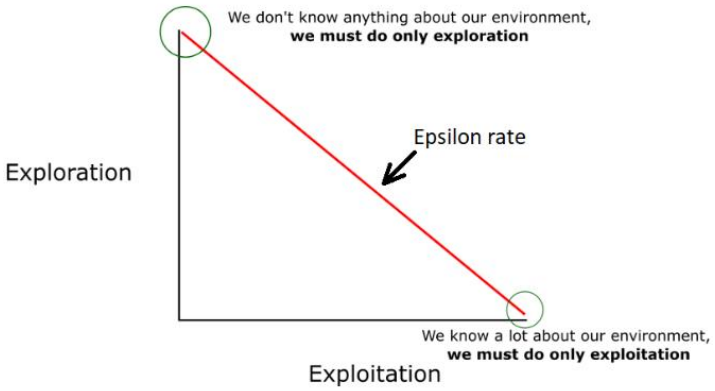


Εικόνα 7. Τρόπος επιλογής ενέργειας με την παράμετρο epsilon .

Όσο ο πράκτορας εξοικειώνεται με το περιβάλλον , θα πρέπει να αρχίσει να επαναλαμβάνει από τις πράξεις που έχει ήδη ανακαλύψει, αυτές που έχουν υψηλότερες τιμές Q, ενώ ταυτόχρονα εξερευνεί λιγότερο καταστάσεις που δεν έχει δει .

Στην πράξη, αυτό επιτυγχάνεται θέτοντας ένα αρχικό epsilon σε μια τιμή που αποφασίζει ο προγραμματιστής, η οποία μειώνεται σταδιακά με την πάροδο του χρόνου. Έτσι, όσο ο πράκτορας αλληλοεπιδρά με το περιβάλλον του, μαθαίνει ποιες ενέργειες έχουν μεγαλύτερη αξία, με αποτέλεσμα οι τιμές που εισάγονται στον πίνακα Q να αρχίζουν να συγκλίνουν στις βέλτιστες τιμές [7].

Η φθορά της τιμής epsilon μπορεί να επιτευχθεί είτε με τη χρήση ενός σταθερού αριθμού που αφαιρείται σε κάθε επανάληψη, είτε με τη χρήση ενός μεταβαλλόμενου αριθμού που επηρεάζεται από μια άλλη μεταβλητή. Ιδανικά η μείωση της παραμέτρου epsilon θα έπρεπε να βασίζεται απευθείας σε τιμές Q που έχουν ήδη ανακαλυφθεί και όχι να μειώνεται σταθερά σε κάθε επανάληψη.

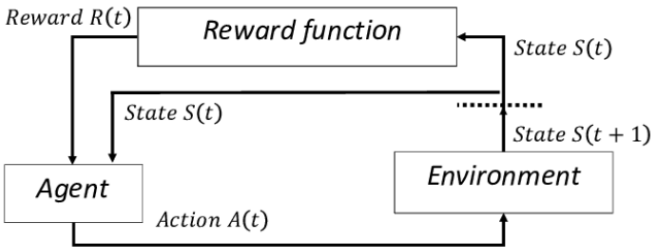


Εικόνα 8. Γραφική αναπαράσταση της μείωσης της υπέρ-παραμέτρου epsilon με σταθερό αριθμό μείωσης.

3.4 Συνάρτηση Ανταμοιβής

Στην ενισχυτική μάθηση ο στόχος του πράκτορα περιγράφεται από ένα σήμα που επιστρέφεται από το περιβάλλον σε αυτόν σε κάθε χρονικό βήμα και ονομάζεται **ανταμοιβή**. Η τιμή της ορίζεται από την αλληλεπίδραση του πράκτορα με το περιβάλλον : όσο «καλύτερη» είναι η ενέργεια που κάνει ο πράκτορας, τόσο μεγαλύτερη είναι και η τιμή της ανταμοιβής. Αυτό δεν σημαίνει απαραίτητα ότι ο πράκτορας στοχεύει στην μεγιστοποίηση της άμεσης ανταμοιβής από την κάθε ενέργεια. Σκοπός του θα έπρεπε να είναι η μεγιστοποίηση της συνολικής ανταμοιβής που μπορεί να αποκτήσει σε βάθος χρόνου, μετά την επιλογή πληθώρας ενεργειών.

Για την επίτευξη της μακροπρόθεσμης ανταμοιβής χρησιμοποιείται μια συνάρτηση μέσω της οποίας το κεντρικό πρόβλημα διασπάται σε υπό-προβλήματα. Ο πράκτορας επιβραβεύεται για την επίλυση του καθενός από αυτά, συνήθως με μία τιμή ανάλογη της δυσκολίας. Η συνάρτηση είναι υπεύθυνη για την καθοδήγηση του πράκτορα ώστε να επιλέξει τις «σωστές» ενέργειες και εν τέλει να φτάσει στην επίλυση του κεντρικού προβλήματος. Ως είσοδο παίρνει την επόμενη κατάσταση και αξιολογεί την ενέργεια που έκανε ο πράκτορας βάσει των αποτελεσμάτων. Έπειτα υπολογίζει την ανταμοιβή και την επιστρέφει στον πράκτορα για τον υπολογισμό της τιμής Q.



Εικόνα 10. Σχεδιάγραμμα για την επίδραση της συνάρτησης ανταμοιβής στην ενισχυτική μάθηση.

Κεφάλαιο 4. Τεχνικές και Μέθοδοι

4.1 Ο αλγόριθμος Epsilon Greedy

Το πρόβλημα του **Multi-Armed Bandit** χρησιμοποιείται στην ενισχυτική μάθηση για να περιγράψει την έννοια της λήψης αποφάσεων σε συνθήκες αβεβαιότητας. Δηλαδή την περίπτωση στην οποία ο πράκτορας επιλέγει μια από k διαφορετικές ενέργειες και λαμβάνει μια ανταμοιβή βάσει της επιλογής του. Για αυτή την επιλογή, υποθέτουμε ότι κάθε πιθανή ενέργεια έχει ξεχωριστή κατανομή ανταμοιβών και υπάρχει τουλάχιστον μία ενέργεια που αποδίδει τη μέγιστη αριθμητική ανταμοιβή. Η κατανομή πιθανοτήτων των ανταμοιβών που αντιστοιχούν σε κάθε ενέργεια είναι διαφορετική και άγνωστη στον πράκτορα. Ο στόχος του είναι να προσδιορίσει ποια να επιλέξει ώστε να λάβει τη μέγιστη ανταμοιβή μετά από ένα σύνολο δοκιμών.

Ο αλγόριθμος εξερεύνησης **Epsilon Greedy** είναι υπεύθυνος για τον έλεγχο του προβλήματος της εναλλαγής μεταξύ εξερεύνησης (**explore**) και εκμετάλλευσης (**exploit**) μιας κατάστασης .Είναι μια στρατηγική που αναλαμβάνει να εκτελέσει μια ενέργεια ανακάλυψης (exploratory action) με πιθανότητα epsilon και μια άπληστη ενέργεια (greedy action) με πιθανότητα 1-epsilon [8] .

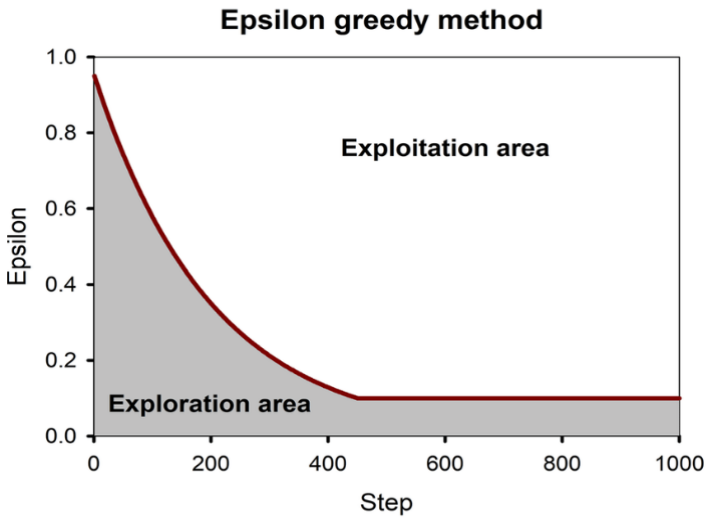
```
p = random()

if p < ε:
   pull random action
else:
   pull current-best action
```

Εικόνα 11. Ψευδοκώδικας για τον αλγόριθμο Epsilon Greedy

Πλεονεκτήματα αλγορίθμου Epsilon-Greedy σε σύγκριση με αντίστοιχους άπληστους αλγορίθμους:

- Εκμεταλλεύεται(exploits) περισσότερο από τους άλλους αλγορίθμους (είναι ο πιο άπληστος).
- Μπορεί να δώσει αποτελέσματα με λιγότερα δείγματα από αυτά που είναι απαραίτητα για αντίστοιχους αλγορίθμους.
- Είναι απλός και εύκολος στην εφαρμογή του.



Εικόνα 12. Παράδειγμα συνάρτησης Epsilon Greedy [9].

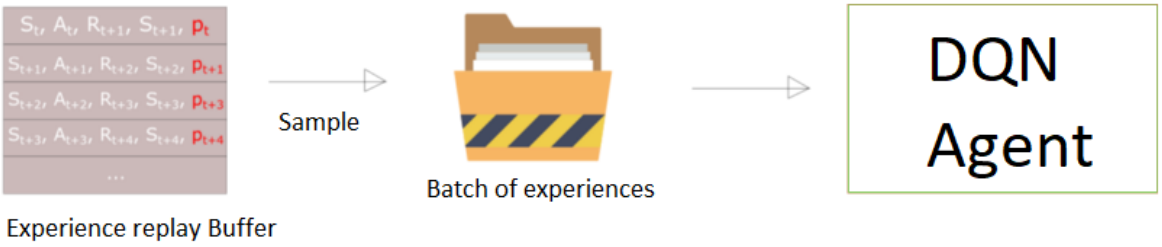
Σε κάθε βήμα ένας τυχαίος αριθμός δημιουργείται από το μοντέλο. Αν ο αριθμός είναι μικρότερος από το epsilon για αυτό το βήμα, το μοντέλο διαλέγει μια τυχαία ενέργεια (explore) , ενώ αν δεν είναι διαλέγει την καλύτερη ενέργεια από αυτές που έχει μάθει(exploit).

4.2 Experience Replay

Μια τεχνική μνήμης που χρησιμοποιείται συχνά μαζί με τον αλγόριθμο του Q-learning είναι αυτή του **Experience Replay**.Στην συγκεκριμένη τεχνική αποθηκεύονται εμπειρίες του πράκτορα σε ένα σταθερού μεγέθους **buffer**, μέσω του οποίου επιτρέπεται η επαναχρησιμοποίηση δεδομένων πολλαπλές φορές για την εκπαίδευση του πράκτορα.

Η επανάληψη της εμπειρίας(experience replay) εφαρμόζεται συνήθως σε ένα κυκλικού τύπου buffer που διαγράφει παλιές εμπειρίες του πράκτορα για να δώσει χώρο ώστε να προστεθούν καινούργιες. Με αυτό τον τρόπο, ο πράκτορας χρησιμοποιεί, ανά σταθερά

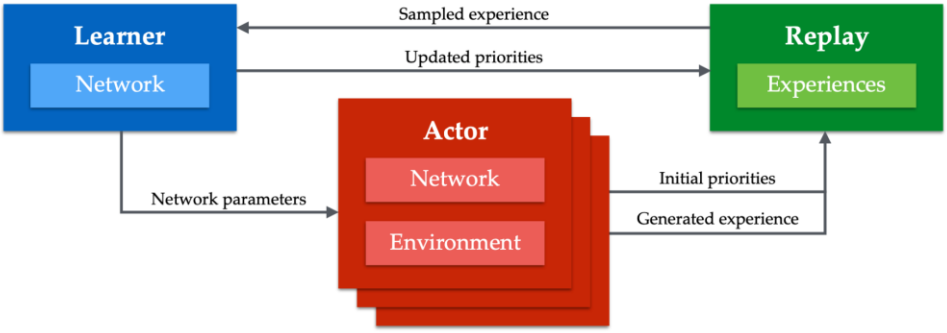
διαστήματα, σύνολα εμπειριών σταθερού μεγέθους (**batch of experiences**) για την εκπαίδευση του, τα οποία στη συνέχεια σβήνονται από τη μνήμη και ελευθερώνουν χώρο για την προσθήκη νέων εμπειριών [10].



Εικόνα 13. Διάγραμμα κατεύθυνσης συνόλου εμπειριών από τον Experience Replay Buffer προς τον πράκτορα.

Η πιο συνηθισμένη στρατηγική δειγματοληψίας που χρησιμοποιείται στην ενισχυτική μάθηση είναι η ομοιόμορφη δειγματοληψία, στην οποία κάθε εμπειρία στο buffer επιλέγεται με ίση πιθανότητα. Μερικές άλλες στρατηγικές δειγματοληψίας, είναι οι εξής:

- Η επανάληψη εμπειρίας προτεραιότητας(**Prioritize Experience Replay**). Αποτελεί μία στρατηγική που δίνει προτεραιότητα στις εμπειρίες που έχουν αναμενόμενα υψηλότερη μαθησιακή πρόοδο. Αυτή η ιεράρχηση των εμπειριών μπορεί να οδηγήσει σε απώλεια ποικιλομορφίας, η οποία όμως μετριάζεται με τη στοχαστική προτεραιότητα, δηλαδή με το να δίνει ο πράκτορας προτεραιότητα σε τυχαίες ενέργειες που θα οδηγήσουν σε αβέβαια αποτελέσματα [11].
- Η επανάληψη κατανεμημένης εμπειρίας προτεραιότητας (**Distributed Prioritized Experience Replay**). Αυτή επιτρέπει σε πολλαπλούς πράκτορες να μαθαίνουν αποτελεσματικά σε μεγαλύτερες τάξεις μεγέθους περισσότερα δεδομένα από ότι είναι δυνατό με τις προηγούμενες μεθόδους, καθώς οι πράκτορες συσσωρεύουν τις εμπειρίες που συλλέξαν από τις αλληλεπιδράσεις με το περιβάλλον σε μια κοινόχρηστη μνήμη επανάληψης [12].



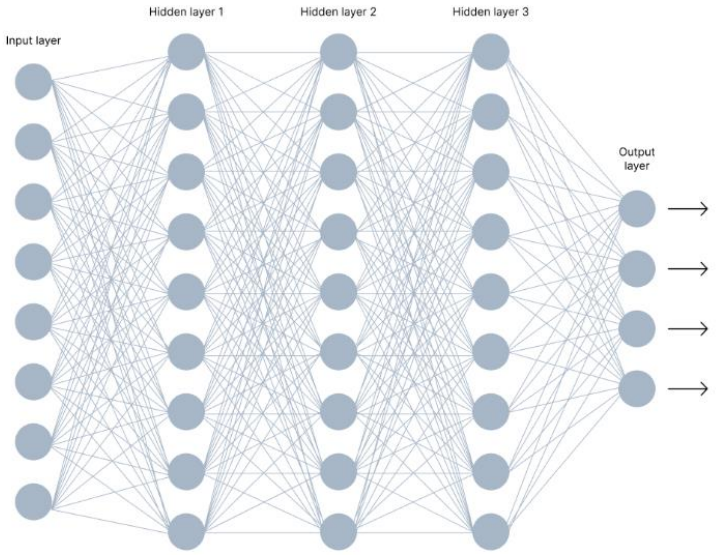
Εικόνα 14.Πολλοί πράκτορες αλληλοεπιδρούν με το περιβάλλον και παράγουν εμπειρίες οι οποίες αποθηκεύονται σε μια κοινόχρηστη μνήμη επανάληψης .Ο κάθε πράκτορας ξεχωριστά παίρνει ένα υποσύνολο των εμπειριών από τη μνήμη και το χρησιμοποιεί για την εκπαίδευση του.

Κεφάλαιο 5. Βαθιά Νευρωνικά Δίκτυα στην Ενισχυτική Μάθηση

Τα νευρωνικά δίκτυα έχουν αποτελέσει ένα απαραίτητο εργαλείο για ένα ευρύ φάσμα εφαρμογών όπως η ταξινόμηση εικόνων, η αναγνώριση ομιλίας και η επεξεργασία φυσικής γλώσσας. Έχει αποδειχθεί πως πετυχαίνουν εξαιρετική προγνωστική ακρίβεια και σε πολλές περιπτώσεις όμοια με αυτή της ανθρώπινης απόδοσης. Η βαθιά ενισχυτική μάθηση συνδυάζει τα τεχνητά νευρωνικά δίκτυα σε ένα πλαίσιο που βοηθάει τον πράκτορα να μάθει πως να πετυχαίνει τους στόχους του. Χρησιμοποιεί τεχνικές εκτίμησης σφάλματος προσέγγισης συναρτήσεων, βελτιστοποίησης στόχων(target optimization) και αντιστοιχίζει καταστάσεις και ενέργειες με τις ανταμοιβές που προκύπτουν [13].

5.1 Τρόπος Λειτουργίας Βαθιών Νευρωνικών Δικτύων

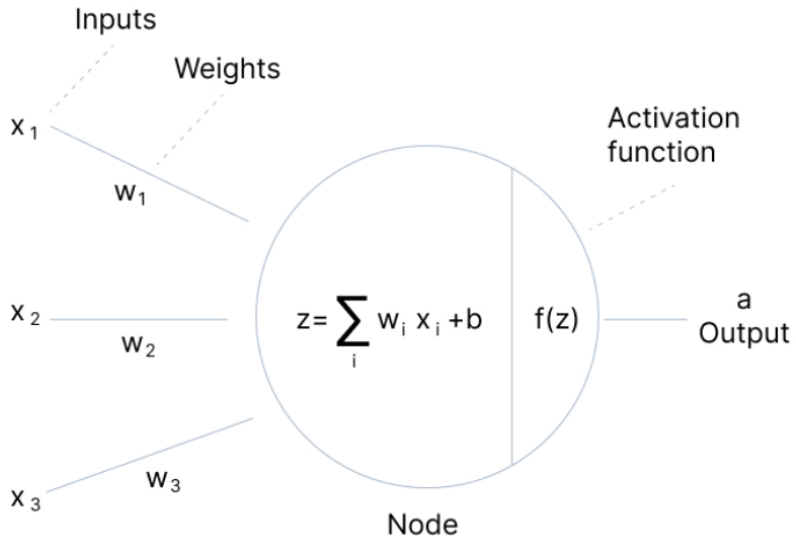
Ένα Βαθύ Νευρωνικό Δίκτυο (DNN) είναι μια συλλογή νευρώνων οργανωμένων σε μια ακολουθία πολλαπλών στρωμάτων. Οι νευρώνες λαμβάνουν ως είσοδο τις ενεργοποιήσεις των νευρώνων από το προηγούμενο στρώμα και εκτελούν έναν απλό υπολογισμό(συνήθως ένα άθροισμα εισόδων επηρεασμένο από κάποιο βάρος που ακολουθείται από μια μη-γραμμική ενεργοποίηση). Οι νευρώνες του δικτύου υλοποιούν από κοινού μια σύνθετη, μη-γραμμική χαρτογράφηση από την είσοδο στην έξοδο . Αυτή η χαρτογράφηση χρησιμοποιεί τα δεδομένα για να προσαρμόσει κατάλληλα τα βάρη κάθε νευρώνα με τέτοιο τρόπο ώστε κάθε πιθανή είσοδος να οδηγεί στη σωστή έξοδο.



Εικόνα 15. Παράδειγμα από Βαθύ Νευρωνικό Δίκτυο

5.2 Συναρτήσεις Ενεργοποίησης

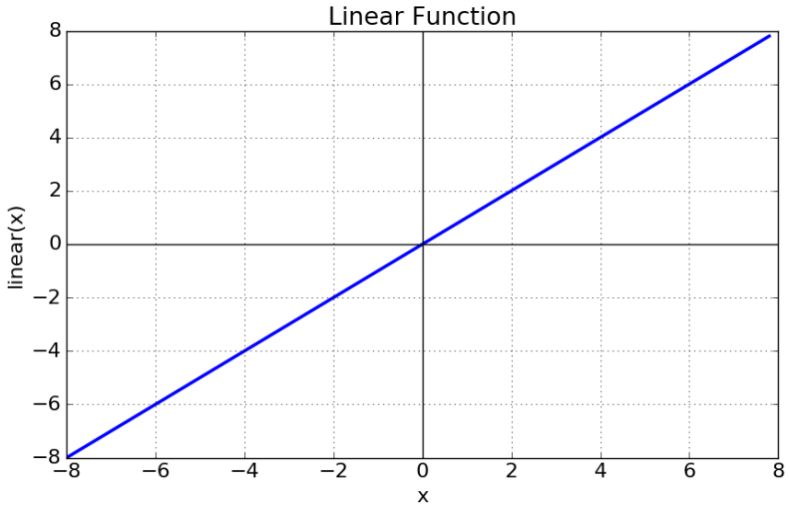
Ο κύριος ρόλος των συναρτήσεων ενεργοποίησης είναι να μετατρέψουν μέσα στο νευρώνα το άθροισμα των εισόδων (όπως έχει διαμορφωθεί από τα βάρη) σε μια τιμή εξόδου που θα τροφοδοτήσει το επόμενο κρυφό στρώμα ή θα οδηγήσει σε έξοδο του νευρωνικού δικτύου.



Εικόνα 16 . Παράδειγμα υπολογισμού της εξόδου του νευρώνα. Πιο αναλυτικά γίνεται άθροισμα των εισόδων που έχουν πολλαπλασιαστεί από τα βάρη , προσθήκη της πόλωσης και ύστερα χρησιμοποίηση αυτού του αποτελέσματος σαν είσοδο στη συνάρτηση ενεργοποίησης f.

Γραμμική Συνάρτηση Ενεργοποίησης

Η συνάρτηση γραμμικής ενεργοποίησης (**linear activation function**) γνωστή και ως "συνάρτηση καμίας ενεργοποίησης" είναι η συνάρτηση όπου η έξοδος είναι ανάλογη της τιμής της εισόδου. Η συνάρτηση δεν επηρεάζει καθόλου την είσοδο που της δόθηκε, απλά επιστρέφει την έξοδο. Δεν επιλύει το πρόβλημα της πολυπλοκότητας του νευρωνικού δικτύου[14].

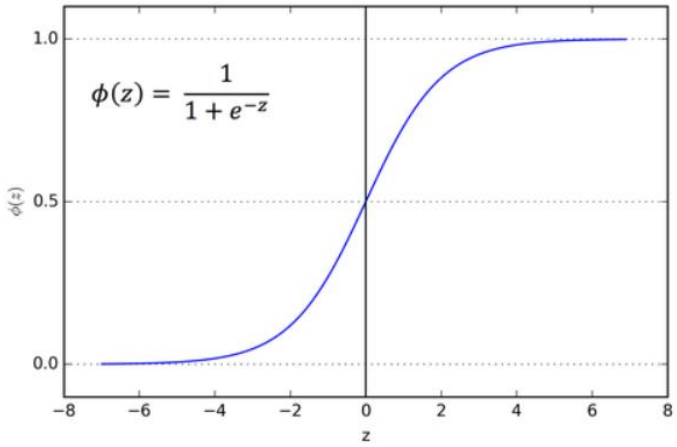


Εικόνα 16. Γραφική αναπαράσταση της Γραμμικής Συνάρτησης Ενεργοποίησης.

Σιγμοειδής Συνάρτηση Ενεργοποίησης

Ο κύριος λόγος χρήσης της Σιγμοειδής Συνάρτησης Ενεργοποίησης σε μερικά μοντέλα είναι η πρόβλεψη της πιθανότητας ως αποτέλεσμα εξόδου. Για αυτό το λόγο η έξοδος της συνάρτησης κυμαίνεται από το 0 έως το 1 και σχηματίζει μια καμπύλη που μοιάζει με το αγγλικό γράμμα "S". Η συνάρτηση είναι διαφοροποιήσιμη, κάτι που σημαίνει ότι μπορεί να ανευρεθεί η κλίση της καμπύλης της σε οποιαδήποτε δύο σημεία.

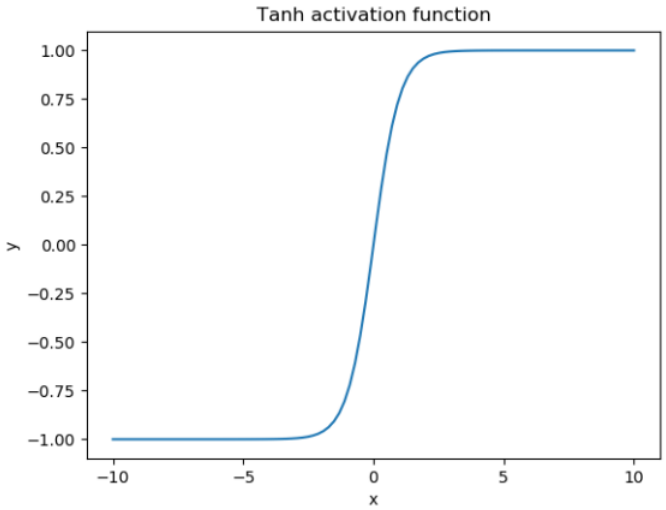
Η Σιγμοειδής Συνάρτηση Ενεργοποίησης μπορεί προκαλέσει το νευρωνικό δίκτυο να «κολλήσει» κατά τη διάρκεια της εκπαίδευσης. Στην περίπτωση, δηλαδή, που εισαχθεί μια πολύ αρνητική τιμή σαν είσοδος, η συνάρτηση θα δώσει μια τιμή εξόδου που βρίσκεται πολύ κοντά στο μηδέν. Λόγω αυτής της συμπεριφοράς , η ενημέρωση των βαρών μπορεί να γίνει αργή και λιγότερο τακτική, με αποτέλεσμα να παρουσιάζονται αμυδρές αλλαγές στο νευρωνικό δίκτυο [15].



Εικόνα 17. Γραφική αναπαράσταση της Σιγμοειδής Συνάρτησης Ενεργοποίησης.

Συνάρτηση Ενεργοποίησης Υπερβολικής Εφαπτομένης (tanh)

Η συνάρτηση ενεργοποίησης tanh αποτελεί μία βελτιωμένη σιγμοειδή συνάρτηση ενεργοποίησης .Το εύρος των τιμών της είναι από το -1 μέχρι το 1 και σχηματίζει μια καμπύλη που μοιάζει με το αγγλικό γράμμα “S”. Το πλεονέκτημα της συνάρτησης υπερβολικής εφαπτομένης έναντι της απλής σιγμοειδούς είναι ότι οι αρνητικές είσοδοι αντιστοιχίζονται σε πολύ αρνητικές εξόδους, οι οποίες όμως δεν βρίσκονται πολύ κοντά στο μηδέν όπως στην σιγμοειδή. Η διαφορά αυτή δίνει μεγαλύτερη απόδοση στην εκπαίδευση των νευρωνικών δικτύων και για αυτό το λόγο η tanh επιλέγεται κατά κύριο λόγο περισσότερο από την προαναφερθείσα[16] .



Εικόνα 18. Γραφική αναπαράσταση της Συνάρτησης Υπερβολικής Εφαπτομένης.

Παρατηρούμε ότι είναι μια μετατοπισμένη και “τεντωμένη” στα άκρα εκδοχή της σιγμοειδούς συνάρτησης.

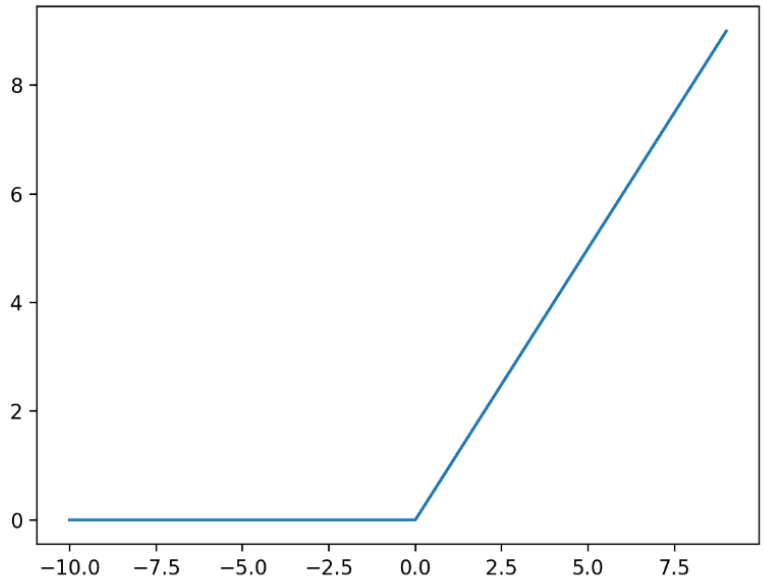
Συνάρτηση Ενεργοποίησης RELU (Rectified Linear Unit)

Η συνάρτηση γραμμικής ενεργοποίησης relu έχει γίνει η πιο διαδομένη συνάρτηση ενεργοποίησης στα νευρωνικά δίκτυα, καθώς το μοντέλο που χρησιμοποιεί καθιστά το δίκτυο πιο εύκολο να εκπαιδευτεί και να επιτυγχάνει συχνά καλύτερη απόδοση.

Πιο συγκεκριμένα ο λόγος επιλογής της συνάρτησης relu ενάντια στις άλλες συναρτήσεις ενεργοποίησης είναι η ανάγκη για μια συνάρτηση που μοιάζει και λειτουργεί σαν γραμμική, ενώ παράλληλα να μπορεί να επιτρέπει την εκμάθηση πολύπλοκων σχέσεων στα δεδομένα, όπως ακριβώς μια μη-γραμμική συνάρτηση.

Τα πλεονεκτήματα της συνάρτησης ενεργοποίησης RELU είναι τα εξής [17]:

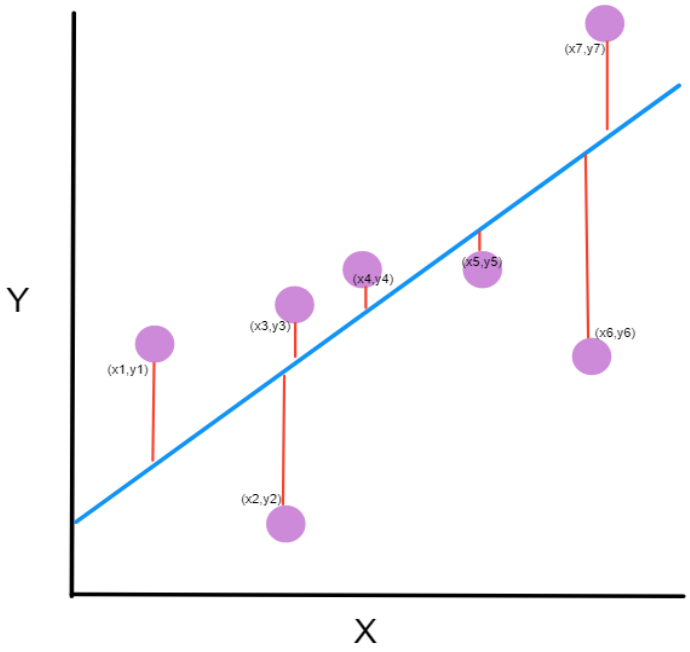
- Υπολογιστική Απλότητα: Είναι απλή στην υλοποίησή της και δεν απαιτεί εκθετικό υπολογισμό όπως η σιγμοειδής ή η tanh.
- Γραμμική Συμπεριφορά: Ένα νευρωνικό δίκτυο είναι πιο εύκολο να βελτιστοποιήσει το αποτέλεσμά της εκπαίδευσης όταν η συμπεριφορά του είναι κοντά στη γραμμική.
- Αντιπροσωπευτικές έξοδοι στις ακραίες τιμές εισόδων: Σε αντίθεση με τη σιγμοειδή και την συνάρτηση υπερβολικής εφαπτομένης, η relu μπορεί να δώσει μηδενική τιμή σε έξοδο.



Εικόνα 19. Γραφική αναπαράσταση της RELU για θετικές και αρνητικές τιμές.

5.3 Συναρτήσεις Εκτίμησης Σφάλματος Προσέγγισης

Οι συναρτήσεις εκτίμησης σφάλματος αξιοποιούνται για την εκτίμηση της ακρίβειας του μοντέλου μηχανικής μάθησης. Στην ουσία αποτελεί μια μέθοδο με την οποία ο προγραμματιστής μπορεί να αξιολογήσει την ποιότητα εκμάθησης του μοντέλου ή, στην προκειμένη περίπτωση, του πράκτορα. Όσο περισσότερο αποκλίνουν τα αποτελέσματα της μάθησης από τα επιθυμητά, τόσο μεγαλύτερο κατά απόλυτη τιμή είναι το αποτέλεσμα της συνάρτησης εκτίμησης σφάλματος. Αντίθετα, αν το μοντέλο έχει την επιθυμητή συμπεριφορά, το αποτέλεσμα της συνάρτησης θα είναι ένας μικρότερος αριθμός. Όσο μεταβάλλονται τα επιμέρους τμήματα του αλγορίθμου με σκοπό την βελτίωση του μοντέλου, η συνάρτηση εκτίμησης σφάλματος μας οδηγεί στην κατασκευή του βέλτιστου αλγορίθμου [18] .



Εικόνα 20 .Παράδειγμα συνάρτηση εκτίμησης σφάλματος προσέγγισης.

Το μέγεθος της απόκλισης του σφάλματος απεικονίζεται με την κόκκινη γραμμή που ενώνει τις τελείες με την ευθεία.

Όσο πιο κοντά στην μπλε ευθεία είναι οι τελείες τόσο μικρότερο σφάλμα εκτίμησης υπάρχει.

Μερικές από τις πιο γνωστές συναρτήσεις εκτίμησης σφάλματος είναι οι παρακάτω:

Συνάρτηση μέσου τετραγωνικού σφάλματος (Mean Square Error):

Το μέσο τετραγωνικό σφάλμα (MSE) είναι το πιο διαδεδομένο καθώς είναι εύκολο να εφαρμοστεί και να κατανοηθεί, ενώ παράλληλα λειτουργεί αρκετά καλά. Στην ουσία δηλώνει πόσο απέχει μια γραμμή παλινδρόμησης από ένα σύνολο σημείων. Αυτό γίνεται παίρνοντας τις αποστάσεις που έχουν τα σημεία από τη γραμμή παλινδρόμησης (οι αποστάσεις είναι τα «λάθη») και τετραγωνίζοντάς τις. Ο τετραγωνισμός είναι απαραίτητος για να αφαιρεθούν τυχόν αρνητικά πρόσημα. Ονομάζεται μέσο τετραγωνικό σφάλμα καθώς υπολογίζει τον μέσο όρο ενός συνόλου σφαλμάτων. Όσο χαμηλότερο είναι το αποτέλεσμα του MSE, τόσο καλύτερη είναι η πρόβλεψη[19].

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Εικόνα 21. Η συνάρτηση μέσου τετραγωνικού σφάλματος

Όπου: y_i είναι η πραγματική τιμή και $\hat{y}_i$ είναι η προβλεπόμενη τιμή.

Συνάρτηση απόλυτου σφάλματος (Mean Absolute Error):

Ορίζουμε το απόλυτο σφάλμα (MAE) ως τον μέσο όρο των απόλυτων διαφορών μεταξύ της πραγματικής και της προβλεπόμενης τιμής. Είναι η δεύτερη πιο συχνά χρησιμοποιούμενη συνάρτηση σφάλματος εκτίμησης σε παλινδρόμηση και μετράει το μέσο μέγεθος των σφαλμάτων, χωρίς να υπολογίζει τις κατευθύνσεις τους. Κάθε σφάλμα πρόβλεψης είναι η διαφορά μεταξύ της πραγματικής τιμής και της προβλεπόμενης τιμής[20].

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

Εικόνα 22. Η συνάρτηση απόλυτου τετραγωνικού σφάλματος.

Όπου y_i είναι η πραγματική τιμή και x_i η προβλεπόμενη τιμή.

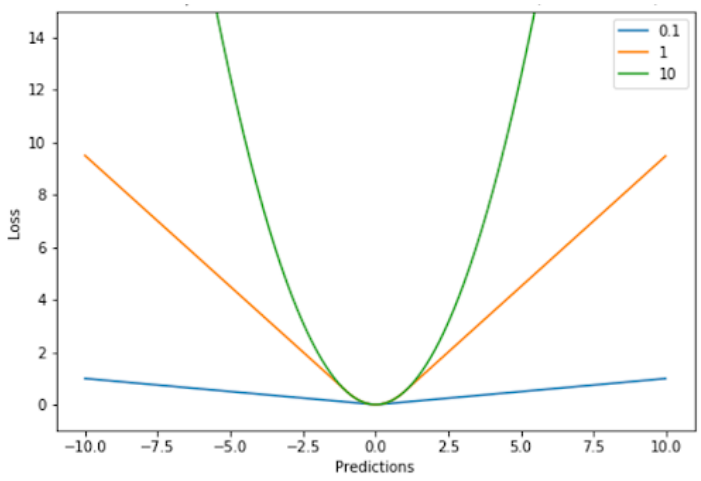
Huber Loss:

Η συνάρτηση εκτίμησης σφάλματος Huber ορίζεται ως ο συνδυασμός των συναρτήσεων μέσου τετραγωνικού σφάλματος(MSE) και μέσου απολύτου σφάλματος (MAE). Η **παράμετρος δ** είναι αυτή που καθορίζει αν θα είναι το σφάλμα μέσο τετραγωνικό ή μέσο απόλυτο. Πιο αναλυτικά, όταν η παράμετρος δ προσεγγίζει το 0 γίνεται χρήση του MSE και όταν προσεγγίζει το ∞ γίνεται χρήση του MAE.

Η επιλογή του δ είναι πολύ κρίσιμη καθώς καθορίζει ποιο σημείο θα θεωρηθεί ως ακραία τιμή (outlier). Ως εκ τούτου, η συνάρτηση εκτίμησης σφάλματος Huber θα μπορούσε να είναι λιγότερο ευαίσθητη σε ακραίες τιμές από τη συνάρτηση απώλειας MSE, ανάλογα με την τιμή της υπερπαραμέτρου. Επομένως η χρήση της προτείνεται κυρίως σε περιπτώσεις που τα δεδομένα είναι επιρρεπή σε ακραίες τιμές(outliers)[21].

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \le \delta, \\ \delta\,|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

Εικόνα 23. Συνάρτηση Huber Loss.



Εικόνα 24.Γραφική Αναπαράσταση των 3 συναρτήσεων που αναφέρθηκαν.

Η πράσινη γραφική παράσταση είναι η συνάρτηση μέσου τετραγωνικού σφάλματος,

Η πορτοκαλί είναι η συνάρτηση Huber loss και η μπλε είναι η συνάρτηση μέσου απολύτου σφάλαματος.

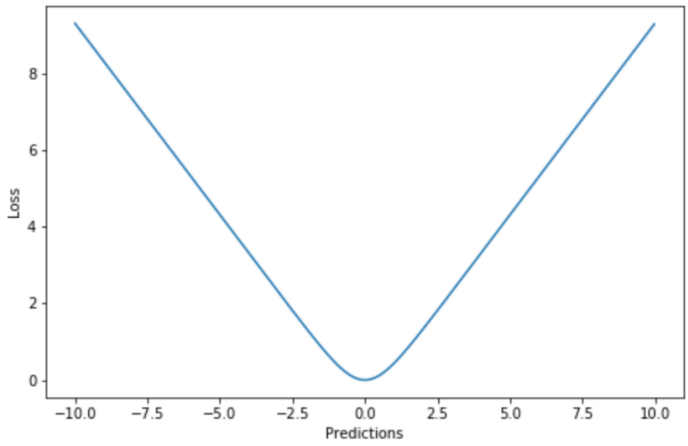
Log-Cosh Loss:

Η συνάρτηση σφάλματος εκτίμησης log-cosh ορίζεται ως ο λογάριθμος του υπερβολικού συνημιτόνου του σφάλματος πρόβλεψης και χρησιμοποιείται κυρίως σε αλγορίθμους μηχανικής μάθησης που βασίζονται σε δέντρα, όπως παραδείγματος χάριν

τον αλγόριθμο XGBoost. Αποτελεί μια συνάρτηση που χρησιμοποιείται σε περιπτώσεις παλινδρόμησης και είναι πολύ πιο ομαλή από την συνάρτηση μέσου τετραγωνικού σφάλματος MSE. Έχει όλα τα πλεονεκτήματα της συνάρτησης Huber ενώ, σε αντίθεση με αυτή, είναι δύο φορές διαφοροποιήσιμη σε όλο το μήκος της[22].

$$L(y, y^p) = \sum_{i=1}^{n} \log(\cosh(y_i^p - y_i))$$

Εικόνα 25.Υπολογισμός συνάρτησης Log-Cosh Loss.



Εικόνα 26. Γραφική απεικόνιση της συνάρτησης Log Cosh Loss.

5.4 Αλγόριθμοι βελτιστοποίησης στην εκπαίδευση Βαθέος Νευρωνικού Δικτύου

Οι αλγόριθμοι βελτιστοποίησης είναι μέθοδοι που χρησιμοποιούνται για την ενημέρωση των βαρών του νευρωνικού δικτύου, με σκοπό τη μείωση των σφαλμάτων προσέγγισης.

Οι πιο συχνοί βελτιστοποιητές (**optimizers**) που χρησιμοποιούνται σήμερα είναι οι παρακάτω :

- Gradient Descent:

  Ο Gradient descent είναι ένας επαναληπτικός αλγόριθμος βελτιστοποίησης για την εύρεση του τοπικού ελάχιστου μιας συνάρτησης που ενημερώνει τα βάρη.

Για να βρεθεί το τοπικό ελάχιστο της συνάρτησης, πρέπει να γίνουν βήματα ανάλογα με το αρνητικό της διαβάθμισης – δηλαδή απομακρυνόμενα από τη διαβάθμιση- της συνάρτησης στο τρέχον σημείο.

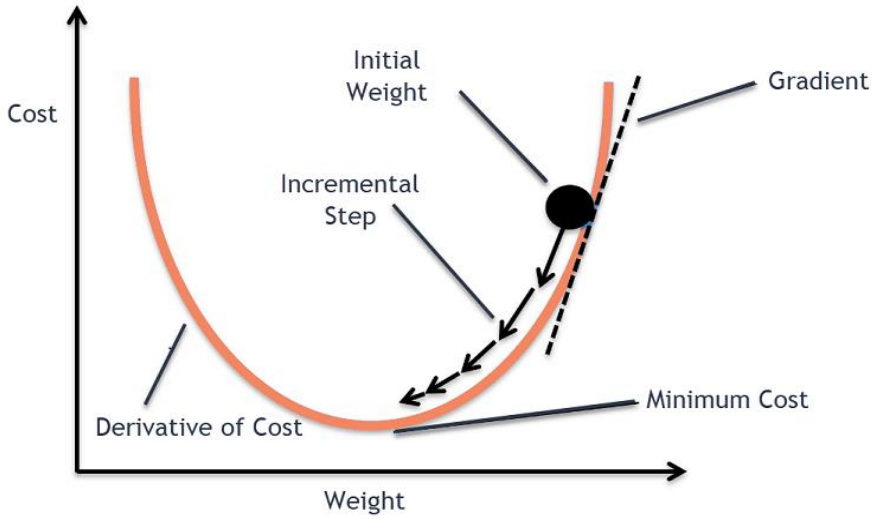
Στην αντίθετη περίπτωση, αν γίνουν βήματα ανάλογα με το θετικό της διαβάθμισης -προχωρώντας στην κλίση προς τα πάνω- προσεγγίζεται το τοπικό μέγιστο της συνάρτησης και στην προκειμένη περίπτωση η διαδικασία ονομάζεται Gradient Ascent .

Η διαδικασία της ελαχιστοποίησης της συνάρτησης για ένα συγκεκριμένο σημείο αποτελείται από δύο βήματα, τα οποία γίνονται επαναληπτικά το ένα μετά το άλλο.

1. Υπολογισμός της κλίσης, δηλαδή της παραγώγου πρώτης τάξης της συνάρτησης σε αυτό το σημείο.
2. Βηματισμός προς την αντίθετη κατεύθυνση από αυτή της αύξησης της διαβάθμισης (gradient) στο σημείο αυτό, κατά a(ρυθμός μάθησης) επί την κλίση στο σημείο[23].

Η διαδικασία τερματίζεται όταν συμβεί ένα από τα παρακάτω:

- Έχει επιτευχθεί ο μέγιστος αριθμός επαναλήψεων.
- Το μέγεθος του βήματος γίνει αμελητέο (λόγω μικρής κλίσης).

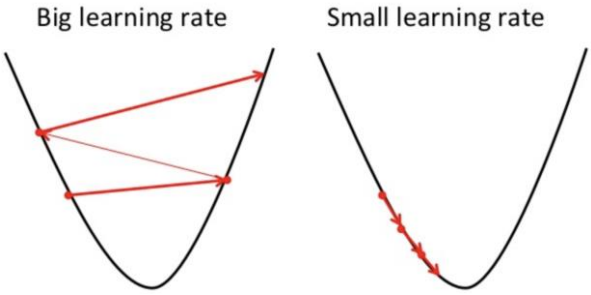


Εικόνα 28. Απεικόνιση του επαναληπτικού αλγορίθμου βελτιστοποίησης Gradient Descent.

$$W_{new} = W_{old} - \alpha * \frac{\partial(Loss)}{\partial(W_{old})}$$

Εικόνα 27. Κανόνας μεταβολής των βαρών

Ο ρυθμός μάθησης a (learing rate) πρέπει να επιλέγεται προσεκτικά ώστε να μπορέσει να εντοπίσει ο Gradient Descent το τοπικό ελάχιστο. Εάν ο ρυθμός μάθησης είναι πολύ υψηλός, μπορεί ο αλγόριθμος να υπερπηδήσει τα ελάχιστα και να συνεχίσει να τα παραλείπει, χωρίς να φτάσει ποτέ σε τοπικό ελάχιστο. Αντίθετα, αν είναι πολύ μικρός, η διαδικασία της εκπαίδευσης μπορεί να γίνει αρκετά χρονοβόρα.



Εικόνα 29. Στην πρώτη περίπτωση το learning rate είναι πολύ μεγάλο με αποτέλεσμα να μη συγκλίνει στο τοπικό ελάχιστο.
Στη δεύτερη περίπτωση το learning rate είναι πολύ μικρό και θα πάρει πολύ χρόνο μέχρι εν τέλει να φτάσει στο τοπικό ελάχιστο.

Η συνάρτηση κόστους μπορεί να έχει πάνω από ένα ελάχιστα. Το ελάχιστο στο οποίο θα καταλήξει ο αλγόριθμος εξαρτάται από το σημείο εκκίνησης και από την τιμή του ρυθμού εκμάθησης. Επομένως η ίδια συνάρτηση, ανάλογα με αυτές τις δύο παραμέτρους, μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα.

<u>Θετικά χαρακτηριστικά του αλγορίθμου Gradient Descent</u> σε σύγκριση με τους υπόλοιπους αλγορίθμους βελτιστοποίησης:

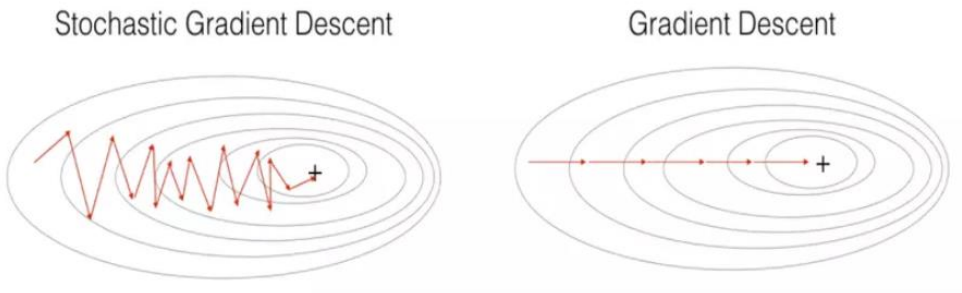
- Υπολογιστικά αποτελεσματικό, καθώς απαιτούνται ενημερώσεις μετά το τέλος μιας εποχής της εκπαίδευσης.
- Ακολουθεί μια σχετικά άμεση πορεία προς το ελάχιστο.
- Εύκολος στην κατανόηση .
- Εύκολος στην υλοποίηση.

Αρνητικά χαρακτηριστικά του αλγορίθμου Gradient Descent σε σύγκριση με τους υπόλοιπους αλγορίθμους βελτιστοποίησης:

- Επειδή η μέθοδος υπολογίζει τη διαβάθμιση(gradient) για ολόκληρο το σύνολο των δεδομένων σε μια ενημέρωση, ο υπολογισμός είναι πολύ αργός.
- Απαιτεί μεγάλη μνήμη και είναι υπολογιστικά ακριβό.
- Μπορεί εύκολα να συγκλίνει σε τοπικά ελάχιστα τα οποία απέχουν πολύ από το ολικό ελάχιστο[24].

- Stochastic Gradient Descent:

  Η στοχαστική Gradient Descent (συχνά αναφερόμενη ως SGD) είναι μια επαναληπτική μέθοδος για τη βελτιστοποίηση μιας αντικειμενικής συνάρτησης με κατάλληλες ιδιότητες ομαλότητας (π.χ. διαφοροποιήσιμη ή υποδιαφορίσιμη). Μπορεί να θεωρηθεί ως μια στοχαστική προσέγγιση του αλγορίθμου Gradient Descent, καθώς αντικαθιστά την πραγματική διαβάθμιση(gradient) (υπολογισμένη από **ολόκληρο το σύνολο δεδομένων**) με μια εκτίμηση αυτής (υπολογισμένη από ένα **τυχαία επιλεγμένο υποσύνολο δεδομένων**).
  Ειδικά σε προβλήματα βελτιστοποίησης πολλών διαστάσεων, αυτό μειώνει τον υψηλό υπολογιστικό φόρτο και χρόνο, επιτυγχάνοντας ταχύτερες επαναλήψεις[25].



Εικόνα 30. Απεικόνιση τη διαφοράς του βηματισμού στο χώρο μεταξύ του Stohastic Gradient Descent και του Gradient Descent.
Όπως φαίνεται και από την εικόνα ο Stohastic Gradient Descent ανακαλύπτει περισσότερο το χώρο.

- Adaptive Gradient Descent (AdaGrad):

  Ο περιορισμός του Gradient Descent να έχει σταθερό μέγεθος βήματος (λόγω σταθερής τιμής μάθησης – learning rate) μπορεί να δημιουργήσει πρόβλημα σε συναρτήσεις που έχουν διαφορετική κλίση καμπυλότητας σε διαφορετικές

διαστάσεις και με τη σειρά τους απαιτούν ένα βήμα διαφορετικού μεγέθους σε ένα νέο σημείο.

Ο αλγόριθμος Adaptive Gradient Descent είναι μια παραλλαγή του αλγόριθμου βελτιστοποίησης Gradient Descent που επιτρέπει στο μέγεθος βήματος να προσαρμόζεται αυτόματα με βάση τις κλίσεις της καμπύλης που εμφανίζονται κατά τη διάρκεια της πορείας της αναζήτησης.

Αυτό επιτυγχάνεται υπολογίζοντας πρώτα ένα μέγεθος βήματος για μια δεδομένη διάσταση και στη συνέχεια χρησιμοποιώντας το υπολογισμένο μέγεθος βήματος για να κάνει μια κίνηση στη διάσταση αυτή, χρησιμοποιώντας τη μερική παράγωγο. Στη συνέχεια, αυτή η διαδικασία επαναλαμβάνεται για κάθε διάσταση στο χώρο αναζήτησης.

Ο αλγόριθμος απαιτεί να οριστεί ένα αρχικό μέγεθος βήματος για όλες τις μεταβλητές εισόδου ως συνήθως, όπως 0.1 ή 0.001 ή παρόμοια. Το βασικό πλεονέκτημα του αλγορίθμου είναι ότι δεν είναι τόσο ευαίσθητος στον αρχικό ρυθμό εκμάθησης όπως άλλοι αλγόριθμοι σαν τον Gradient Descent [26].

$$W_{new} = W_{old} + \frac{\alpha}{\sqrt{cache_{new}} + \epsilon} * \frac{\partial(Loss)}{\partial(W_{old})}$$

Εικόνα 31.Μεταβολή βαρών βάσει του αλγορίθμου Adaptive Gradient Descent.

- Root Mean Square Propagation (RMS-Prop):

  Το Root Mean Square Propagation είναι μια παραλλαγή του AdaGrad στην οποία ο ρυθμός εκμάθησης είναι ένας εκθετικός μέσος όρος των κλίσεων των σημείων, αντί για το άθροισμα των τετραγωνικών κλίσεων τους. Στην ουσία ο RMS-Prop είναι ένας συνδυασμός του AdaGrad και του Gradient Descent with momentum. Ο ρυθμός εκμάθησης προσαρμόζεται αυτόματα και διαφοροποιείται για τις εκάστοτε παραμέτρους[27].

$$cache_{new} = \gamma * cache_{old} + (1 - \gamma) * (\frac{\partial(Loss)}{\partial(W_{old})})^2$$

Εικόνα 32. Συνάρτηση υπολογισμού Root Mean Square Propagation.

- Adaptive Moment Estimation (Adam)

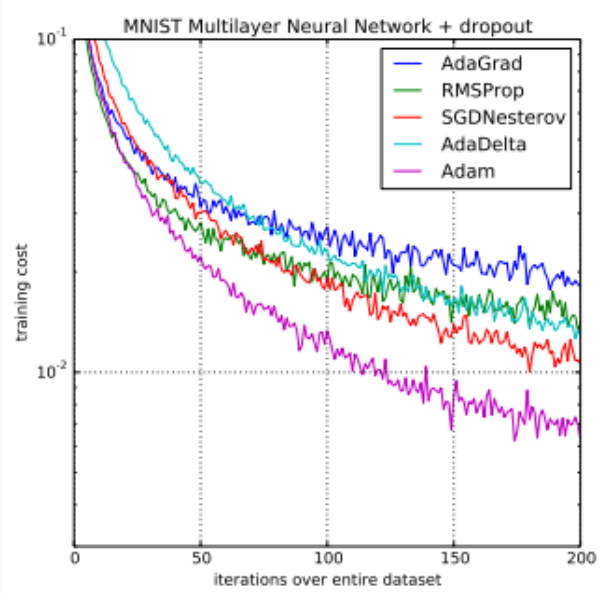
  Ο αλγόριθμος βελτιστοποίησης Adam είναι μια παραλλαγή του Stohastic Gradient Descent που έχει πρόσφατα γίνει γνωστός ευρύτερα για εφαρμογές βαθιάς μάθησης στην υπολογιστική όραση και την επεξεργασία φυσικής γλώσσας.

  Όπως προαναφέρθηκε, ο SGD διατηρεί έναν ενιαίο ρυθμό μάθησης (που ονομάζεται a) για όλες τις ενημερώσεις βάρους, ο οποίος δεν αλλάζει κατά τη διάρκεια της εκπαίδευσης.

  Ο αλγόριθμος Adam συνδυάζει τα πλεονεκτήματα δύο άλλων παραλλαγών του SGD , του αλγορίθμου Adaptive Gradient Algorithm και του Root Mean Square Propagation που αναφέρθηκαν προηγουμένως[28].

  Τα επικρατέστερα χαρακτηριστικά του Adam είναι τα εξής:

  - Το μέγεθος βήματος και κατά συνέπεια ο ρυθμός μάθησης που λαμβάνεται από τον Adam σε κάθε επανάληψη οριοθετείται κατά προσέγγιση με μια υπερπαράμετρο μεγέθους βήματος.
  - Το μέγεθος του βήματος είναι ανεξάρτητο από το μέγεθος της κλίσης, με αποτέλεσμα να αποφεύγονται μικρά τοπικά ελάχιστα και saddle points.
  - Ο Adam σχεδιάστηκε για να συνδυάζει τα πλεονεκτήματα αφενός του Adagrad -που λειτουργεί καλά με αραιά δεδομένα- και αφετέρου του RMSprop -που λειτουργεί καλά σε on-line τροποποιήσεις του ρυθμού μάθησης-.



Εικόνα 33. Σύγκριση του Adam με άλλους αλγορίθμους βελτιστοποίησης για την εκπαίδευση ενός νευρωνικού δικτύου .Παρατηρούμε ότι έχει το μικρότερο υπολογιστικό κόστος .

Πέρα από τα πλεονεκτήματα του αλγορίθμου, αξίζει να σημειωθεί πως σε μερικές περιπτώσεις ο Adam δεν συγκλίνει σε μια βέλτιστη λύση για ορισμένες περιοχές. Δεδομένου αυτού του ελαττώματος, δίνεται κίνητρο για περαιτέρω χρήση και μελέτη μιας νέας παραλλαγής του αλγορίθμου με όνομα AMSGrad ο οποίος παρέχει λύση στα προβλήματα σύγκλισης και οδηγεί σε βελτιωμένες επιδόσεις [29].

Κεφάλαιο 6. Υλοποίηση αλγορίθμων στα δύο σενάρια της εργασίας

6.1 One Hot Encoding

Τα κατηγορικά δεδομένα (categorical data) αναφέρονται σε μεταβλητές που χαρακτηρίζονται από τιμές που έχουν τη μορφή λέξεων. Ορισμένοι αλγόριθμοι μηχανικής μάθησης μπορούν να λειτουργήσουν απευθείας με κατηγορικά δεδομένα, ανάλογα πάντα με την υλοποίηση (π.χ. ένα δέντρο αποφάσεων). Ωστόσο οι περισσότεροι προϋποθέτουν οι μεταβλητές εισόδου ή εξόδου που χρησιμοποιούνται να είναι κάποιος πραγματικός αριθμός. Για το λόγο αυτό απαιτείται οποιαδήποτε κατηγορικά δεδομένα να αντιστοιχίζονται σε ακέραιους αριθμούς.

Το One Hot Encoding είναι μια μέθοδος μετατροπής δεδομένων για την προετοιμασία τους σε έναν αλγόριθμο για καλύτερη πρόβλεψη. Με τη μέθοδο αυτή κάθε τιμή κατηγορίας μετατρέπεται σε μια νέα κατηγορική στήλη και εκχωρείται μια δυαδική τιμή 0 ή 1 στις στήλες αυτές. Έτσι λοιπόν, κάθε κατηγορική τιμή αναπαρίσταται ως δυαδικό διάνυσμα. Οι τιμές στις στήλες που έχουν δημιουργηθεί με One Hot Encoding είναι αρχικοποιημένες στο μηδέν και η αντίστοιχη κατηγορηματική τιμή που είναι ενεργή κάθε στιγμή μετατρέπεται στον πίνακα από 0 σε 1[30] .

Στην τρέχουσα εργασία, στην περίπτωση του παίκτη εναντίον αντιπάλου, η μέθοδος One Hot Encoding χρησιμοποιείται για να συμβολίσει ποια ομάδα έχει την κατοχή της μπάλας. Έτσι στην περίπτωση που την κατοχή την έχει η αριστερή ομάδα -δηλαδή ο πράκτορας- στον πίνακα που απεικονίζει τον χώρο καταστάσεων μπαίνουν οι τιμές

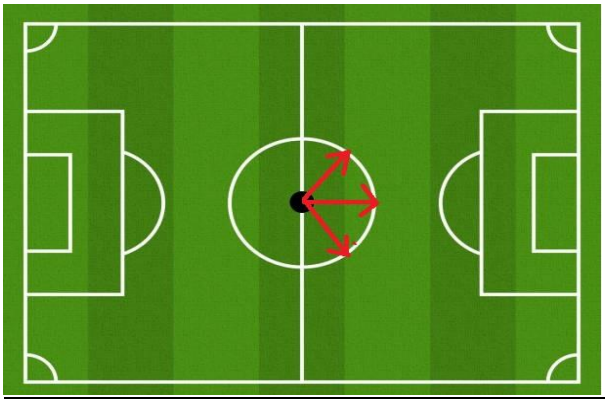
(1,0,0), στην περίπτωση που την μπάλα δεν την έχει κανένας παίκτης μπαίνουν οι τιμές (0,1,0) και στην περίπτωση που την κατοχή την έχει η δεξιά ομάδα -δηλαδή ο αντίπαλος- μπαίνουν οι τιμές (0,0,1).

6.2 Συναρτήσεις Ανταμοιβής στα δύο σενάρια παιχνιδιού

Συνάρτηση ανταμοιβής στο σενάριο του παίκτη χωρίς αντίπαλο:

Η διαδικασία που πρέπει να ακολουθήσει ο πράκτορας προκειμένου να φτάσει στη μεγάλη περιοχή και να βάλει goal είναι ιδιαίτερα απλή.

Εφόσον ο ίδιος ξεκινάει από το κέντρο και κινείται προς την πλευρά του αντίπαλου τέρματος (πάνω δεξιά , κάτω δεξιά και ευθεία δεξιά), το μόνο που πρέπει να προσέξει είναι μη φύγει πέρα από τα άνω και κάτω περιθώρια, από τα οποία αν εκτελέσει σουτ η μπάλα θα βγει εκτός ορίων. Επομένως η συνάρτηση ανταμοιβής που θα κατασκευαστεί θα είναι εξίσου απλή με το συγκεκριμένο πρόβλημα.



Εικόνα 34. Αναπαράσταση των πιθανών κινήσεων του παίκτη στο χώρο.
Η μαύρη κουκίδα αντιπροσωπεύει την αρχική θέση του παίκτη και τα κόκκινα βέλη αντιπροσωπεύουν τις πιθανές κινήσεις του .

Η πρώτη συνάρτηση επιβράβευσης που δοκιμάστηκε ήταν απλώς μια αύξηση του reward στην περίπτωση που ο πράκτορας έβαζε goal και αντίστοιχα μια μείωση του reward στην περίπτωση που ο παίκτης έβγαζε την μπάλα εκτός ορίων.

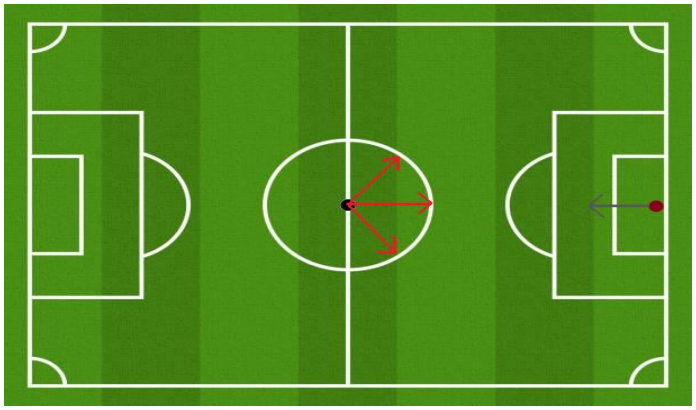
Ενώ η παραπάνω συνάρτηση είχε καλά αποτελέσματα, παρατηρήθηκε ότι πράκτορας ακόμα και στο τέλος της εκπαίδευσης έκανε πολλά βήματα που δεν είχαν κάποια χρησιμότητα για την επίλυση του προβλήματος.

Επομένως για την αποφυγή αυτών των άχρηστων πράξεων(actions), δημιουργήθηκε η επόμενη και τελική συνάρτηση ανταμοιβής, η οποία αποθαρρύνει τον πράκτορα από το να κάνει πράξεις που δεν τον βοηθούν άμεσα στο να βάλει goal. Αυτή περιλαμβάνει τα εξής κριτήρια:

- Αύξηση της ανταμοιβής κατά 10 μονάδες αν ο παίκτης βάλει goal στο αντίπαλο τέρμα.
- Μείωση της ανταμοιβής κατά 10 μονάδες αν η μπάλα βγει εκτός των ορίων του γηπέδου.
- Μείωση της ανταμοιβής κατά 10 μονάδες αν τελειώσει το επεισόδιο και δεν έχει μπει goal.
- Αύξηση της ανταμοιβής κατά 2 μονάδες αν ο παίκτης έχει την μπάλα και κάνει βήμα στην ίδια ευθεία με αυτή του αντίπαλου τέρματος.
- Η ανταμοιβή μειώνεται ανάλογα με την ευκλείδεια απόσταση της μπάλας από το αντίπαλο τέρμα (reward = reward – ($\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$) * 0.3).

<u>Συνάρτηση ανταμοιβής στο σενάριο του παίκτη με αντίπαλο:</u>

Σε αυτή την περίπτωση η διαδρομή που πρέπει να ακολουθήσει ο παίκτης για να σκοράρει είναι αρκετά πιο περίπλοκη και άμεσα εξαρτώμενη από την συμπεριφορά του αντιπάλου. Πιο αναλυτικά, ο πράκτορας ξεκινάει από το κέντρο του γηπέδου, χρειάζεται να προστατέψει την κατοχή της μπάλας και να περάσει τον αντίπαλο παίκτη και σκοπός του είναι να βάλει goal. Ο αντίπαλος από την αρχή του παιχνιδιού κινείται προς την μπάλα και δεν θα σταματήσει μέχρι να την κλέψει και να την κρατήσει συνολικά για πάνω από 5 χρόνο-βήματα, να βγει η μπάλα εκτός των ορίων του γηπέδου, ή να μπει goal.



Εικόνα 35.Αναπαράσταση των πιθανών κινήσεων του πράκτορα και του αντιπάλου παίκτη.
Η σκούρο κόκκινη κουκκίδα αντιπροσωπεύει τον αντίπαλο και το γκρι βέλος την αρχική του κίνηση προς την μπάλα.

Η συνάρτηση επιβράβευσης για το συγκεκριμένο σενάριο ορίστηκε ως εξής για κάθε χρόνο-βήμα:

- Αν ο πράκτορας **σκοράρει** στο αντίπαλο τέρμα αυξάνεται η ανταμοιβή κατά 60 μονάδες.
- Αν **κάνει σουτ** και η μπάλα βγει εκτός ορίων, η ανταμοιβή μειώνεται κατά 10 μονάδες.
- Αν ο πράκτορας **έχει κατοχή,** η ανταμοιβή αυξάνεται κατά 5 μονάδες, ενώ αν αντίθετα κατοχή έχει ο αντίπαλος η ανταμοιβή μειώνεται κατά 5 μονάδες.
- Αν ο πράκτορας **δεν κάνει σουτ** αλλά βγάλει την **μπάλα εκτός ορίων** από την πλευρά του αντιπάλου τέρματος, η ανταμοιβή μειώνεται κατά 20 μονάδες.
- Αν ο **αντίπαλος** παίκτης έχει την **κατοχή της μπάλας** για 1 χρόνο-βήμα, η ανταμοιβή μειώνεται κατά 20 μονάδες, αν την έχει για 2 βήματα κατά 30 μονάδες, για 3 βήματα 40 μονάδες, για 4 και για 5 βήματα 50 μονάδες.
- Αν η **μπάλα περάσει τη γραμμή του κέντρου** με κατεύθυνση προς τα πίσω ή **βγει εκτός ορίων** από πλάγια του γηπέδου, η ανταμοιβή μειώνεται κατά 35 μονάδες.
- Αν ο **αντίπαλος** είναι **μακριά από την μπάλα** και ο πράκτορας έχει κατοχή και κάνει βήματα προς το αντίπαλο τέρμα, η ανταμοιβή αυξάνεται κατά 2 μονάδες σε κάθε βήμα, ενώ αντίθετα αν δεν πηγαίνει προς το αντίπαλο τέρμα η ανταμοιβή μειώνεται κατά 5 μονάδες ανά βήμα.
- Αν ο **πράκτορας έχει κατοχή, είναι απέναντι από τον αντίπαλο** σε μικρή απόσταση και κάνει την **πρώτη φορά ντρίμπλα,** η ανταμοιβή αυξάνεται κατά 20 μονάδες.

- Αν ο **πράκτορας έχει περάσει τον αντίπαλο**, έχει κατοχή και κάνει πρώτη φορά σπριντ η ανταμοιβή αυξάνεται κατά 20 μονάδες.
- Αν ο **πράκτορας κάνει σουτ εντός της μεγάλης περιοχής** η ανταμοιβή αυξάνεται κατά 25 μονάδες.
- Για την αποφυγή πράξεων που δεν θα είχαν χρησιμότητα στη λύση του προβλήματος, όπως και στο σενάριο του παίκτη χωρίς αντίπαλο, προστέθηκε η **συσχέτιση της ανταμοιβή**ς με την **απόσταση της μπάλας από το αντίπαλο τέρμα**(reward = reward – ($\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$) * 0.3).

Δεδομένου της πολύπλοκης συνάρτησης επιβράβευσης που χρησιμοποιήθηκε, όπως θα δούμε και στη συνέχεια, ο πράκτορας εφευρίσκει πολύ γρήγορα τρόπους για να περάσει τον αντίπαλο είτε μέσω ντρίμπλας είτε μέσω σπριντ και καταφέρνει να σκοράρει από τα πρώτα κιόλας επεισόδια.

6.3 Υπέρ-παράμετροι Νευρωνικού Δικτύου και αλγορίθμου

Υπέρ-παράμετροι Νευρωνικού δικτύου για το σενάριο του παίκτη χωρίς αντίπαλο :

Παράμετρος	**Τιμή**
Actors	1
Batch Size	64
Discount Factor (gamma)	0.99
Epsilon	1.0
Epsilon decay	0.0005
Epsilon end	0.01
Learning Rate	0.00115
Optimizer	Adam
Neurons for each Network layer	128
Activation Function	RELU

Υπέρ-παράμετροι Νευρωνικού δικτύου για το σενάριο του παίκτη με αντίπαλο :

Παράμετρος	**Τιμή**
Actors	1
Batch Size	64
Discount Factor (gamma)	0.99
Epsilon	1.0
Epsilon decay	0.0005
Epsilon end	0.01
Learning Rate	0.0011
Optimizer	Adam
Neurons for each Network layer	262
Activation Function	RELU

6.4 Αποτελέσματα Εκπαίδευσης

Η μέθοδος που χρησιμοποιείται για την ανακάλυψη του περιβάλλοντος είναι η epsilon greedy. Όπως είδαμε και παραπάνω, ο πράκτορας κάνει περιστασιακά τυχαίες πράξεις που οδηγούν σε εξερεύνηση του περιβάλλοντος, ενώ όσο εκπαιδεύεται επιλέγει να κάνει όλο και περισσότερο τις πράξεις που δίνουν άμεσα ή έμμεσα τη μέγιστη ανταμοιβή.

Ένα πρόβλημα που προκύπτει από αυτή τη μέθοδο είναι πως στην αρχή της εκπαίδευσης είναι πιθανό να μην γίνει αποτελεσματικά η εξερεύνηση του περιβάλλοντος και ως συνέπεια ο πράκτορας μπορεί να μην δοκιμάσει ποτέ να κάνει τις σωστές ενέργειες για τις αντίστοιχες καταστάσεις.

Το αποτέλεσμα μιας τυχαίας αρχής εκπαίδευσης, σαν και αυτή, είναι ο πράκτορας να επαναλαμβάνει λάθος πράξεις για ορισμένες καταστάσεις και να μην καταλήγει ποτέ στη λύση του συνολικού προβλήματος, ή να καταλήγει σε αυτή σπάνια και τυχαία.

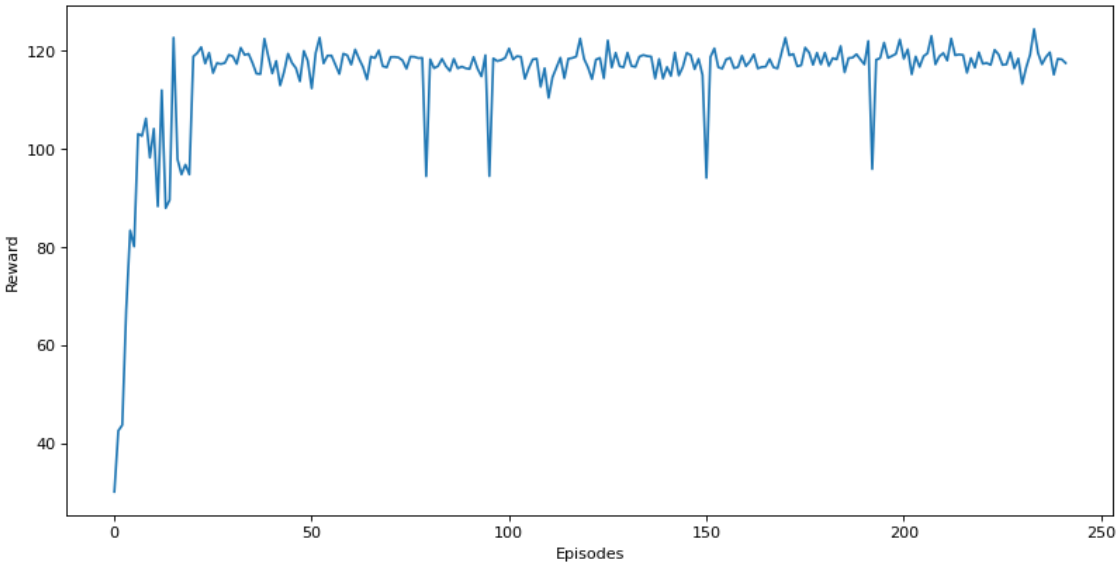
6.4.1 Κακό ξεκίνημα εκπαίδευσης στο περιβάλλον

Στο **σενάριο του παίκτη χωρίς αντίπαλο**, εξαιτίας της έλλειψης αντιπάλου -άρα και τυχαιότητας στο περιβάλλον- ο πράκτορας βρίσκει τη λύση του προβλήματος ακόμα και όταν στα αρχικά στάδια της εκπαίδευσης δεν τυχαίνει να εκτελεί τις καλύτερες δυνατές ενέργειες. Δηλαδή στο τελικό στάδιο της εκπαίδευσης ο παίκτης σκοράρει με ποσοστό επιτυχίας 100% στα τελευταία 50 επεισόδια και επιπλέον κάνει λιγότερα βήματα μέχρι να σκοράρει σε κάθε επεισόδιο.

Παρ΄ όλα αυτά, τα επεισόδια που χρειάζονται για τη λύση του προβλήματος και τη σύγκλιση του αλγορίθμου είναι περισσότερα από αυτά που θα χρειαζόντουσαν στην περίπτωση ενός καλού ξεκινήματος εκπαίδευσης.

Πιο αναλυτικά, αν και στα βήματα του πρώτου επεισοδίου ο πράκτορας δέχεται πολύ χαμηλά rewards, μετά από μερικά επεισόδια καταλήγει να ακολουθεί τις σωστές πράξεις που τον οδηγούν στη λύση του προβλήματος: δηλαδή να σκοράρει στο άδειο τέρμα. Ο πράκτορας έβρισκε και πάλι συνεχώς καλύτερους τρόπους να προσεγγίσει το πρόβλημα και να λάβει τη μέγιστη ανταμοιβή. Η συγκεκριμένη εκπαίδευση, ωστόσο, λόγω της «κακής» αρχής της διήρκησε 250 επεισόδια (110 περισσότερα από το καλό ξεκίνημα εκπαίδευσης) μέχρι τα αποτελέσματα να συγκλίνουν, καθώς στην αρχική κατάσταση ο πράκτορας έκανε περισσότερες άχρηστες πράξεις(actions).

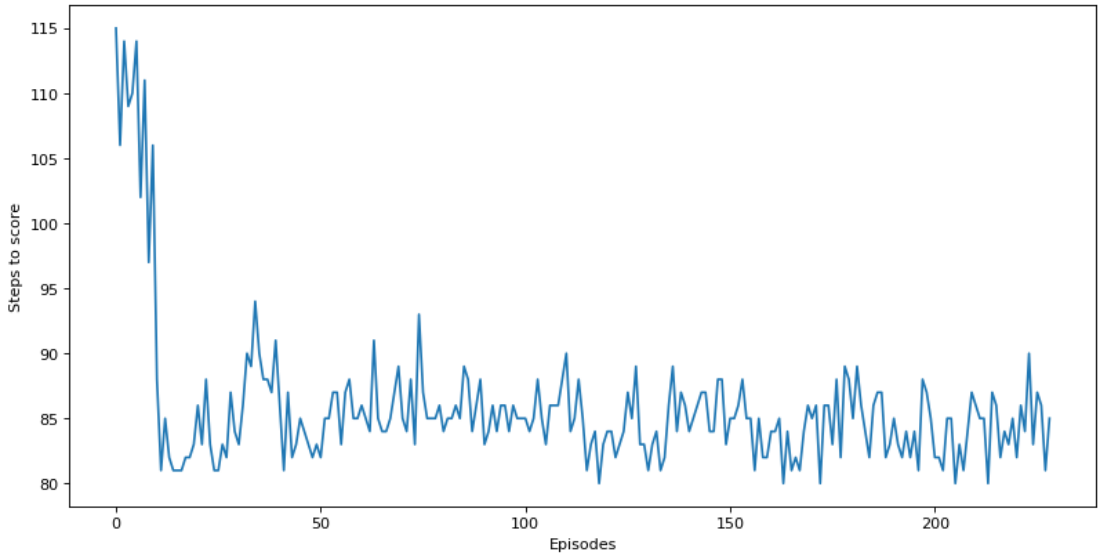
Στη συγκεκριμένη περίπτωση, στη γραφική αναπαράσταση της ανταμοιβής ανά επεισόδιο (Εικόνα 36) παρατηρείται απότομη αύξηση της ανταμοιβής από τα πρώτα κιόλας επεισόδια. Αυτό οφείλεται στην απλότητα του συγκεκριμένου προβλήματος, καθώς και στην απλή, αλλά πολύ αποτελεσματική συνάρτηση ανταμοιβής που χρησιμοποιήθηκε.



Εικόνα 36.Γραφική αναπαράσταση της ανταμοιβής ανά επεισόδιο

στην περίπτωση μια κακής αρχής εκπαίδευσης.

Εξαιτίας της συνάρτησης ανταμοιβής και συγκεκριμένα της συσχέτισης της ανταμοιβής με την απόσταση της μπάλας από το αντίπαλο τέρμα, ο πράκτορας έκανε συνεχώς λιγότερα βήματα μέχρι τη μεγάλη περιοχή όπου έκανε σουτ. Όπως φαίνεται και από το παρακάτω διάγραμμα (Εικόνα 37), το κακό ξεκίνημα της εκπαίδευσης δεν επηρέασε τη συμπεριφορά αυτή. Η απότομη κλίση προς τα κάτω του γραφήματος στα πρώτα επεισόδια της εκπαίδευσης είναι ενδεικτική της αποτελεσματικής συνάρτησης ανταμοιβής.



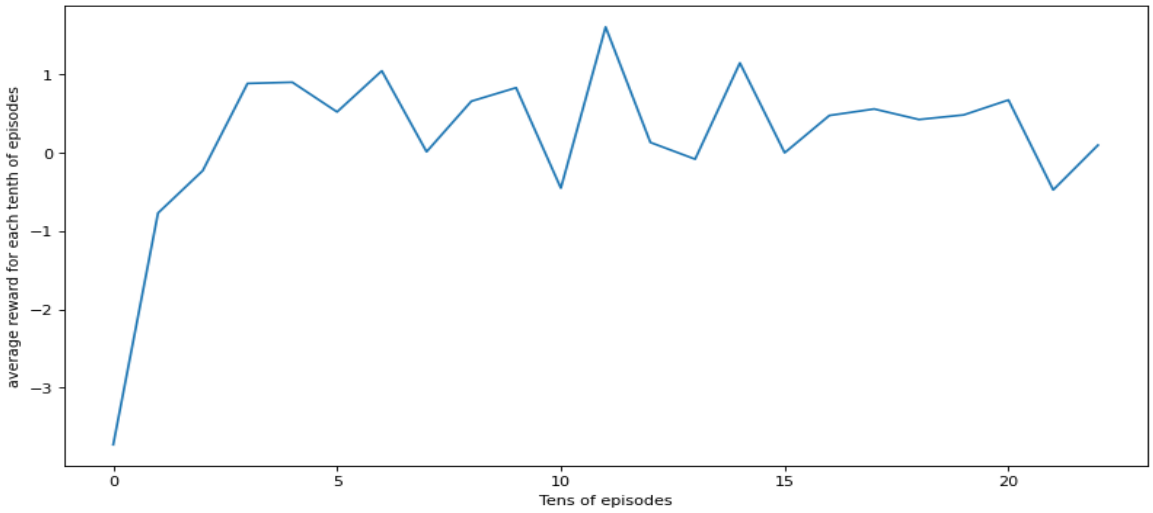
Εικόνα 37. Γραφική αναπαράσταση των βημάτων που χρειάστηκαν μέχρι να σκοράρει

Στην περίπτωση της κακής αρχής εκπαίδευσης.

Στο **<u>σενάριο του παίκτη εναντίον αντιπάλου.</u>** στην περίπτωση ενός κακού ξεκινήματος εκπαίδευσης τα πράγματα είναι πιο ξεκάθαρα. Όταν ο παίκτης δεν μαθαίνει από τα αρχικά στάδια της εκπαίδευσης να περνάει τον αντίπαλο και να έχει κατεύθυνση προς το αντίπαλο τέρμα, καταλήγει να επιλέγει πολύ σπάνια ενέργειες που τον οδηγούν στο να σκοράρει .

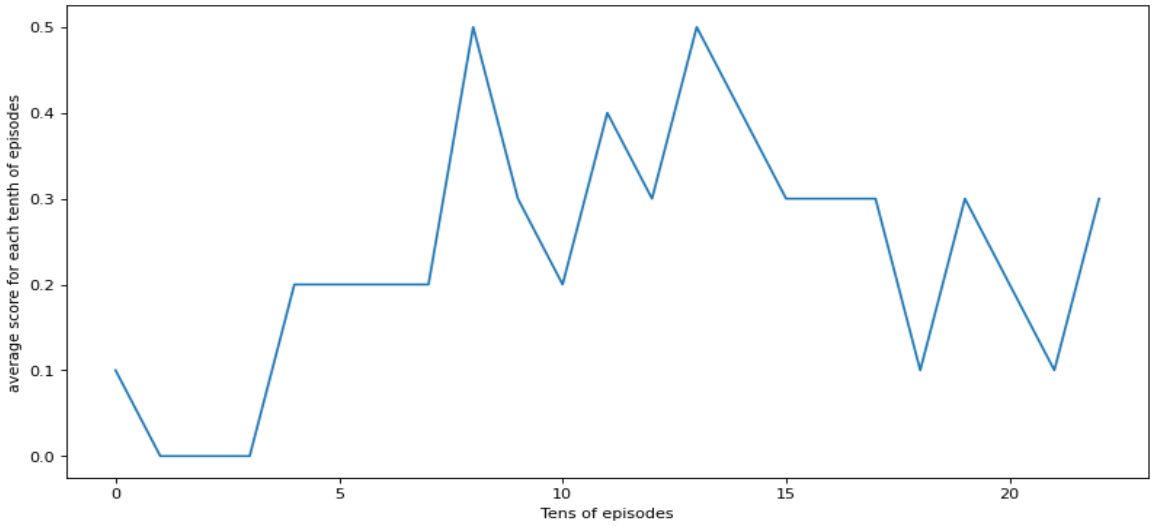
Πιο αναλυτικά, όταν δεν ξεκινάει από την αρχή να εκτελεί πράξεις οι οποίες επιστρέφουν μεγάλη ανταμοιβή από τη συνάρτηση ανταμοιβής , τότε ούτε στο τέλος της εκπαίδευσης ακολουθεί τη διαδοχή των πράξεων που θα οδηγήσουν στη μέγιστη ανταμοιβή και κατά συνέπεια στη λύση του προβλήματος.

Κάτι εξίσου σημαντικό που παρατηρήθηκε στη διάρκεια της εκπαίδευσης είναι ότι αν ο παίκτης δεν επιλέξει να κάνει σουτ από τα πρώτα επεισόδια της εκπαίδευσης, καταλήγει να μαθαίνει να παίζει σε κάθε επεισόδιο χωρίς να κάνει σουτ. Προτιμάει δηλαδή να οδηγεί την μπάλα στο τέρμα για να βάλει goal, παρά να κάνει σουτ από μακριά.



Εικόνα 38.Γραφική αναπαράσταση της μέσης ανταμοιβής για κάθε δεκάδα επεισοδίων στο σενάριο του παίκτη εναντίον αντιπάλου.
Παρατηρούμε ότι το πρώτο επεισόδιο ξεκινάει από ανταμοιβή κοντά στις- 4 μονάδες και στη συνέχεια γίνεται μια απότομη αύξηση στις -1 μονάδες.
Η αύξηση αυτή οφείλεται στην αποδοτική συνάρτηση ανταμοιβής που χρησιμοποιήθηκε.

Επειδή ο πράκτορας μαθαίνει από την αρχή να επιλέγει πράξεις που τον οδηγούν στη λήψη χαμηλής ανταμοιβής και επομένως δεν έχει ανακαλύψει τις πράξεις που οδηγούν στο goal, κατά τη διάρκεια της εκπαίδευσης δεν παρουσιάζονται μεγάλες αλλαγές στα αποτελέσματα του σκορ. Τέλος, η εκπαίδευση σταματά με τον παίκτη να σκοράρει με ποσοστό επιτυχίας μόνο 24% .



Εικόνα 39.Γραφική αναπαράσταση του μέσου όρου goal ανά δεκάδα επεισοδίων.

Η καλύτερη απόδοση που έχει ο πράκτορας είναι με ποσοστό επιτυχίας 50% για 2 επεισόδια

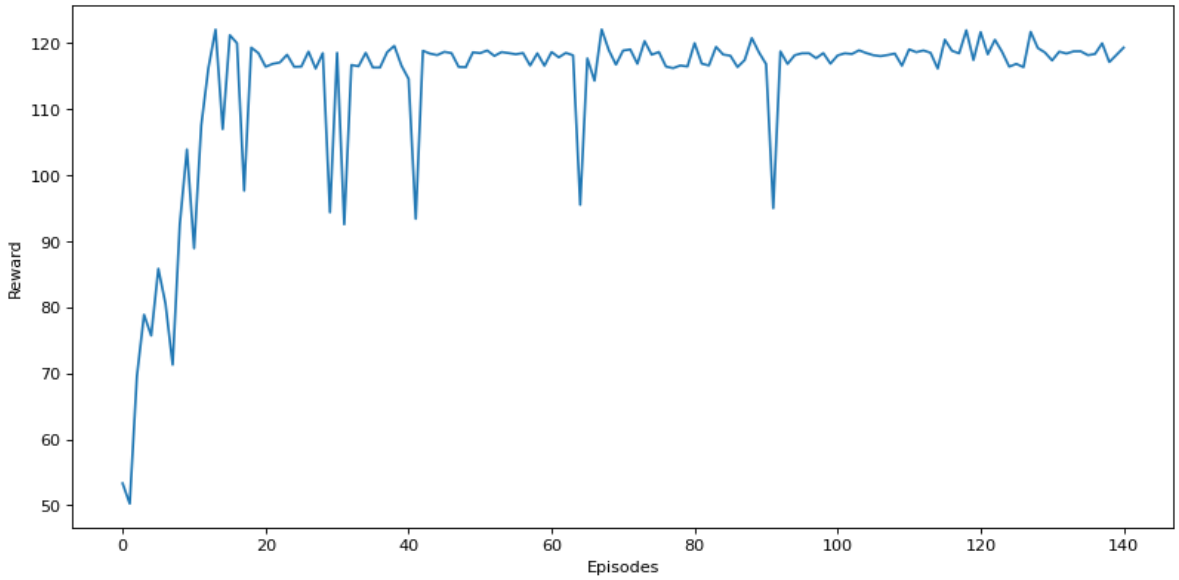
Ενώ στο τέλος της εκπαίδευσης είναι σε θέση να βάζει 24 goal στα 100 επεισόδια.

6.4.2 Καλό ξεκίνημα εκπαίδευσης στο περιβάλλον

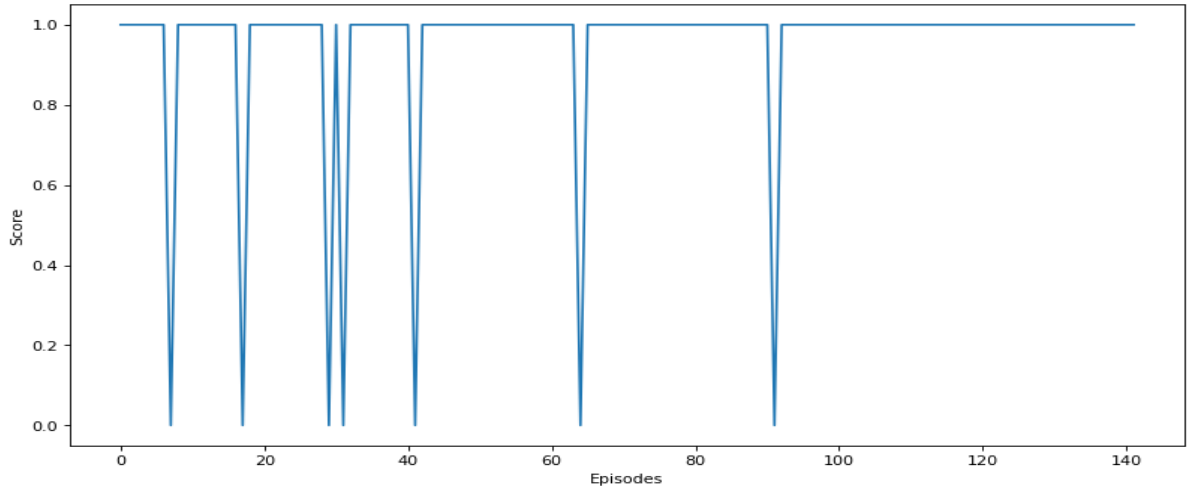
Για **το σενάριο του παίκτη χωρίς αντίπαλο,** στην περίπτωση που ο πράκτορας στα αρχικά στάδια της εκπαίδευσης ανακαλύπτει τις σωστές πράξεις για κάθε δεδομένη κατάσταση, τα αποτελέσματα είναι πολύ καλύτερα. Ο πράκτορας -όπως και στην εκπαίδευση με κακή αρχή- φτάνει να σκοράρει με ποσοστό 100% και τα βήματα μέχρι να μπει goal επίσης μειώνονται.

Η διαφορά με την προηγούμενη προσπάθεια εκπαίδευσης για το ίδιο σενάριο, είναι πως τώρα απαιτούνται πολύ λιγότερα επεισόδια (συγκεκριμένα 110 λιγότερα) για την εύρεση λύσης και για τη σύγκλιση του αλγορίθμου.

Όπως προηγουμένως, έτσι και εδώ παρατηρείται απότομη αύξηση της ανταμοιβής από τα πρώτα κιόλας επεισόδια, καθώς και απότομη μείωση των βημάτων μέχρι να σκοράρει ο παίκτης. Οι δύο παρατηρήσεις αυτές μπορούν να χρησιμοποιηθούν σαν δείκτες ενδεικτικοί της αποτελεσματικότητας της συνάρτησης ανταμοιβής που δημιουργήθηκε.

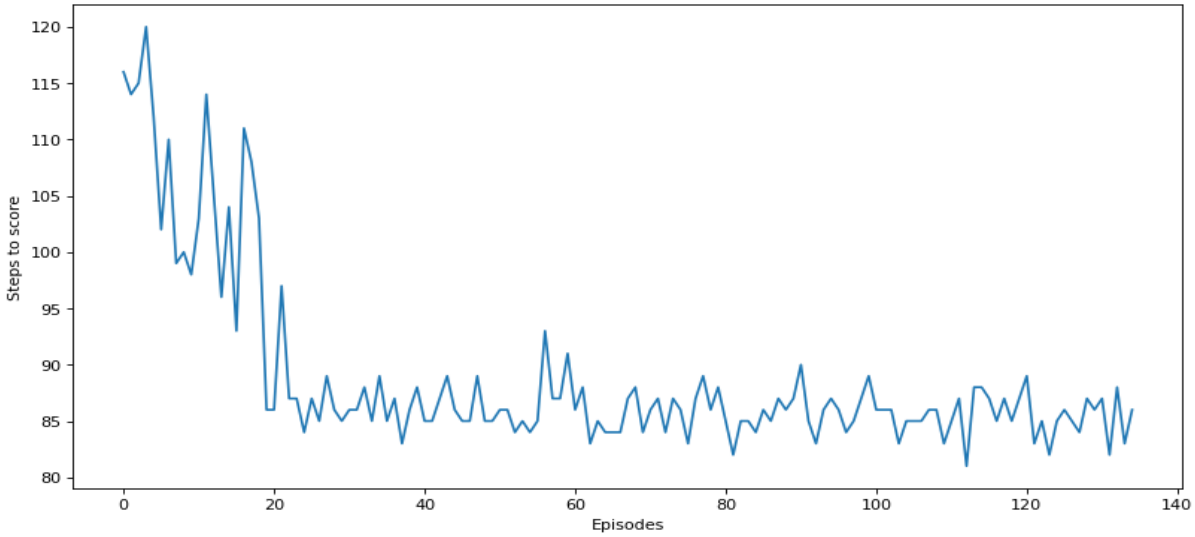


Εικόνα 40. Γραφική αναπαράσταση της επιβράβευσης ανά επεισόδιο για το σενάριο παίκτη χωρίς αντίπαλο περίπτωση ενός καλού ξεκινήματος εκπαίδευσης.



Εικόνα 41. Γραφική αναπαράσταση του σκορ ανά επεισόδιο στο σενάριο του παίκτη χωρίς αντίπαλο.

Παρατηρούμε ότι στα τελευταία 50 επεισόδια σκοράρει με 100% επιτυχία.

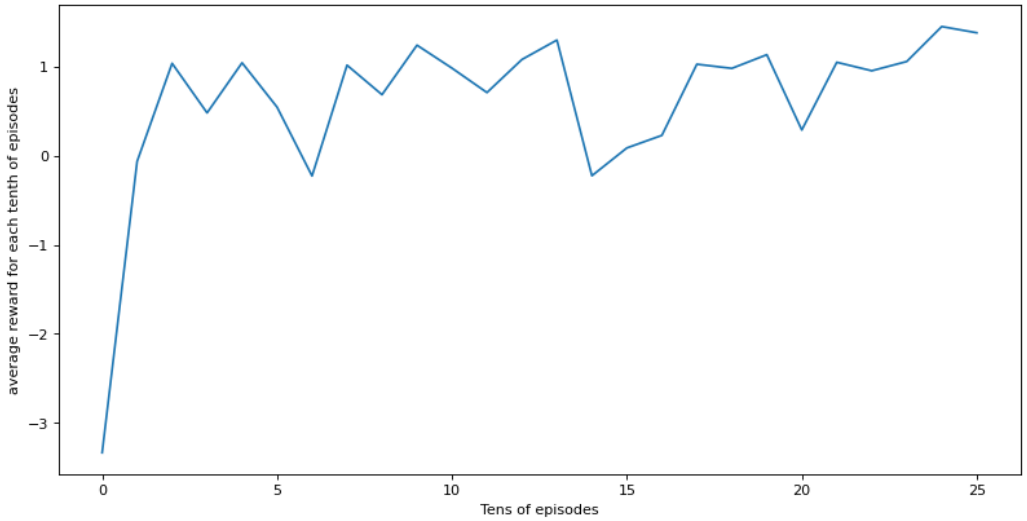


Εικόνα 42 .Γραφική αναπαράσταση των βημάτων που χρειάστηκαν να σκοράρει μέχρι να σκοράρει.

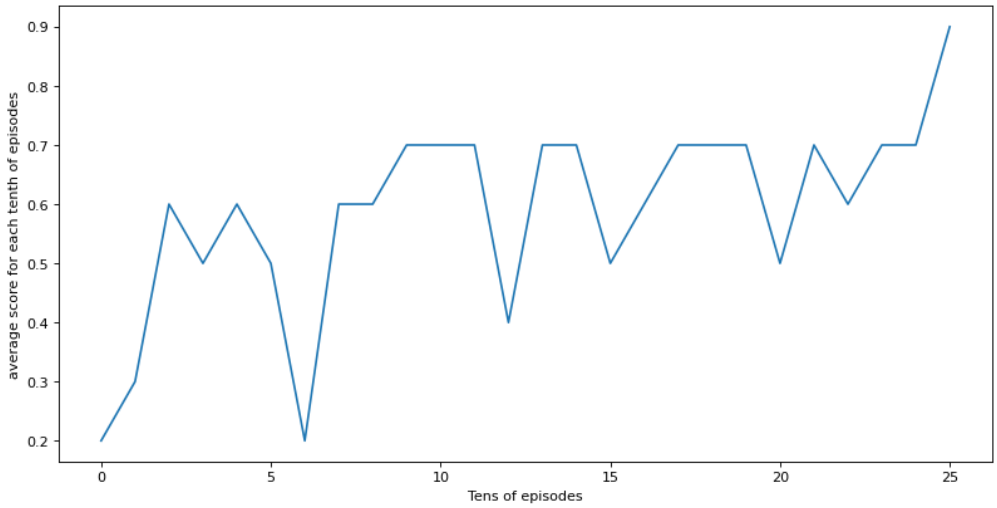
Για το **<u>σενάριο του παίκτη με αντίπαλο,</u>** στην περίπτωση που ο πράκτορας στην αρχή της εκπαίδευσης ανακαλύπτει τις σωστές πράξεις για δεδομένες καταστάσεις στο περιβάλλον, τα αποτελέσματα είναι εμφανώς καλύτερα. Σε αντίθεση με την προηγούμενη εκπαίδευση στο ίδιο σενάριο -όπου ο πράκτορας παρουσίασε ποσοστό επιτυχίας 24%- σε αυτή την περίπτωση καταφέρνει να σκοράρει με ποσοστό επιτυχίας 70%.

Ο λόγος που υπάρχει τόσο μεγάλη διαφορά στο ποσοστό επιτυχίας μεταξύ των δύο αυτών πειραμάτων είναι πως στο δεύτερο ο πράκτορας μαθαίνει, από τα πρώτα κιόλας επεισόδια, να περνάει τον αντίπαλο και να πηγαίνει προς το απέναντι τέρμα. Επίσης σε αυτή την περίπτωση ανακαλύπτει πως είναι χρήσιμο να κάνει ντρίμπλα πριν περάσει τον αντίπαλο και στη συνέχεια σπριντ αφότου είναι σε πλεονεκτική θέση μπροστά του, κατευθυνόμενος προς το απέναντι τέρμα. Ακόμα, ανακαλύπτει πως το σουτ μέσα στη μεγάλη περιοχή είναι προτιμότερο από το να προχωράει με τη μπάλα μέσα στο αντίπαλο τέρμα, καθώς έτσι οι πιθανότητες να του κλέψει τη μπάλα ο αντίπαλος μέσα στη μεγάλη περιοχή είναι μικρότερες.

Συνεπώς στο τέλος της εκπαίδευσης τα επεισόδια στα οποία ο παίκτης δεν σκοράρει είναι εκείνα στα οποία δεν καταφέρνει να περάσει τον αντίπαλο και χάνει την κατοχή της μπάλας, ή εκείνα όπου ο αντίπαλος διώχνει τη μπάλα εκτός των ορίων του γηπέδου.



Εικόνα 43.Γραφική αναπαράσταση της μέσης ανταμοιβής για κάθε δεκάδα επεισοδίων στο σενάριο του παίκτη εναντίον αντιπάλου. Παρ όλο που η ανταμοιβή στο πρώτο επεισόδιο είναι κοντά στις -4 μονάδες, παρατηρούμε ότι γίνεται απότομη αύξηση στα επόμενα επεισόδια καθώς ο πράκτορας μαθαίνει να επιλέγει τις «σωστές» ενέργειες μέσα στο επεισόδιο.

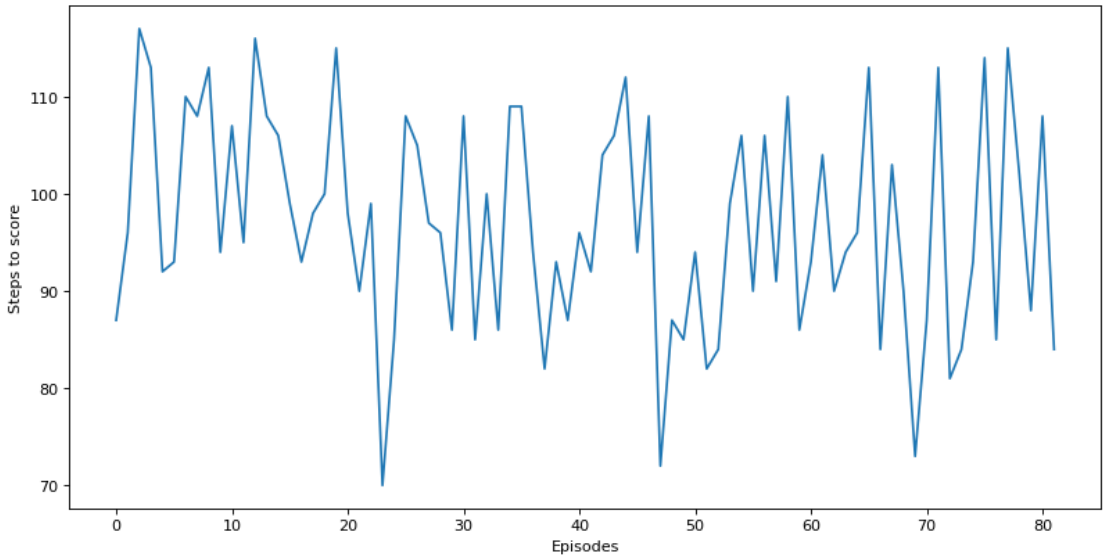


Εικόνα 44. Γραφική αναπαράσταση του μέσου όρου goal για κάθε δεκάδα επεισοδίων. Όπως φαίνεται και από το γράφημα ενώ ο παίκτης ξεκινάει βάζοντας μόνο 2 goal στα πρώτα 10 επεισόδια, πολύ γρήγορα ξεκινάει και βάζει 6 στα 10, ενώ στο τέλος φτάνει να βάζει μέχρι και 9 στα 10 goal.

Στο σενάριο του παίκτη εναντίον αντιπάλου, επειδή οι ενέργειες του πράκτορα είναι άμεσα συνδεδεμένες με την τυχαιότητα του περιβάλλοντος, δηλαδή τις κινήσεις του αντίπαλου παίκτη, σε κάθε επεισόδιο εκτελεί διαφορετικό αριθμό βημάτων.

Αυτό οφείλεται στο ότι η προσέγγιση του αντίπαλου παίκτη δεν είναι πάντα η ίδια, συνεπώς ο πράκτορας μαθαίνει να προσαρμόζεται σε μια σειρά από πιθανές πράξεις και να τις χειρίζεται ανάλογα, ώστε να οδηγηθεί στο να σκοράρει.

Αυτό έχει ως αποτέλεσμα ο αριθμός των βημάτων σε κάθε επεισόδιο να μην ακολουθεί κάποια γραμμική μεταβολή και να μην αποτελεί δείκτη για την ποιότητα της εκπαίδευσης, όπως αντίθετα είναι η μέση ανταμοιβή ανά επεισόδιο και το σκορ ανά επεισόδιο.



Εικόνα 45. Γραφική αναπαράσταση του αριθμού των βημάτων μέχρι ο πράκτορας να σκοράρει για κάθε επεισόδιο.

Παρατηρούμε ότι από το παραπάνω γράφημα δεν μπορούμε να καταλήξουμε σε κάποιο συμπέρασμα.

Κεφάλαιο 7. Απαιτήσεις συστήματος

Η βαθιά ενισχυτική μάθηση αξιοποιεί πολύ έντονα τους πόρους του συστήματος και είναι σε γενικές γραμμές μια υπολογιστικά ακριβή μέθοδος μάθησης. Τα χαρακτηριστικά του συστήματος που χρησιμοποιείται σε κάθε περίπτωση έχουν άμεση σχέση με το χρόνο που χρειάζεται για να ολοκληρωθεί μια εκπαίδευση και ως αποτέλεσμα με τη διευκόλυνση του προγραμματιστή για τη διεξαγωγή πολλών και διαφορετικών πειραμάτων.

Για την παρούσα εργασία το σύστημα που χρησιμοποιήθηκε ήταν το παρακάτω:

Cpu Model	Intel(R) Xeon(R) CPU @ 2.30GHz
Core(s) per socket:	20
Thread(s) per core	2
Ram Memory	36G

Κεφάλαιο 8. Βιβλιοθήκες που χρησιμοποιήθηκαν

Για την εκπαίδευση του πράκτορα και για τα 2 σενάρια στο περιβάλλον του Google Research Football χρησιμοποιήθηκαν οι παρακάτω βιβλιοθήκες της γλώσσας προγραμματισμού Python:

- Gfootball: Δημιουργία περιβάλλοντος και εκτέλεση προσομοίωσης
- Pytorch: Δημιουργία νευρωνικού δικτύου και υλοποίηση υπολογισμών ανάμεσα στους νευρώνες του δικτύου
- Matplotlib: Δημιουργία γραφικών παραστάσεων
- Numpy: Υλοποίηση πράξεων πινάκων καθώς και πράξεων λιστών
- Random: Χρήση για παραγωγή τυχαίων αριθμών
- Math: Υλοποίηση αλγεβρικών πράξεων

Βιβλιογραφία

[1] "Reinforcement Learning: An Introduction", Richard S. Sutton and Andrew G. Barto , 1998.

[2] Stuart J. Russell , Peter Norvig (2010) "Artificial Intelligence: A Modern Approach", Third Edition

[3] "Reinforcement-Learning: State-Of-The-Art ",Marco A.Wiering, Martijn van Otterlo

[4] "Hands-On Q-Learning with Python: Practical Q-learning with OpenAI Gym, Keras, and TensorFlow 1st Edition, Kindle Edition ", Nazia Habib

[5] Chow, Gregory P. (1976). "Analysis and Control of Dynamic Economic Systems"

[6] "Reinforcement Learning: An Introduction", Richard S. Sutton and Andrew G. Barto , 1998.

[7] "An Adaptive Implementation of ε-Greedy in Reinforcement Learning", Alexandredos Santos Mignon ,Ricardo Luisde Azevedo da Rocha

[8] "The Epsilon Greedy Algorithm - a Performance Review ", International Journal of New Technology and Research (IJNTR) ISSN: 2454-4116, Volume-6, Issue-9, September 2020 Pages 01-03,

Riti Agarwal

[9] "Double-deep Q-learning to increase the efficiency of metasurface holograms", Iman Sajedian

[10] "Revisiting Fundamentals of Experience Replay", Published at ICML 2020 ,William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, Will Dabney

[11] "PRIORITIZED EXPERIENCE REPLAY", Published as a conference paper at ICLR 2016 , Tom Schaul, John Quan, Ioannis Antonoglou and David Silver

[12] "DISTRIBUTED PRIORITIZED EXPERIENCE REPLAY" , Published as a conference paper at ICLR 2018 , Dan Horgan , John Quan, David Budden, Gabriel Barth-Maron , Matteo Hessel , Hado van Hasselt , David Silver

[13] "Methods for interpreting and understanding deep neural networks", Digital

Signal, Processing Volume 73, February 2018, GrégoireMontavon, WojciechSamek, Klaus-RobertMüller

[14] "Activation Functions in Neural Networks", V7 Article, July 19 2022,Pragati Baheti

[15] "Sigmoid Activation Function in Selecting the Best Model of Artificial Neural Networks", Journal of Physics: Conference Series, Volume 1471, 1st Bukittinggi International Conference on Education 17-18 October 2019, West Sumatera, Indonesia

[16] "Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks",Tomasz Szandała, October 2020

[17] "Deep Learning (Adaptive Computation and Machine Learning series) ", by Ian Goodfellow, Yoshua Bengio , Aaron Courville , December 12, 2022

[18] "On Loss Functions for Deep Neural Networks in Classification",Katarzyna Janocha, Wojciech Marian Czarnecki

[19] "Mean Squared Error: Love It or Leave It? ",IEEE SIGNAL PROCESSING MAGAZINE [98] JANUARY 2009, Zhou Wang and Alan C. Bovik

[20] "Encyclopedia of Machine Learning",2010,Claude Sammut, Geoffrey I. Webb

[21] "Robust Estimation of a Location Parameter",Huber, Peter J. (1964).

[22] "Statistical Properties of the log-cosh Loss Function Used in Machine Learning",Resve A. Saleh, A.K.Md. Ehsanes Saleh, 12 Aug 2022

[23] "Gradient-based learning applied to document recognition" ,Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Published in: Proceedings of the IEEE (Volume: 86, Issue: 11, November 1998)

[24] "Gradient-Descent-like Ghost Imaging",Wen-Kai Yu,* Chen-Xi Zhu, Ya-Xin Li, Shuo-Fei Wang, and Chong Cao, Published online 2021 Nov 13.

[25] "The Tradeoffs of Large Scale Learning", Bottou, Léon; Bousquet, Olivier (2012)

[26] "Algorithms for Optimization",by Mykel J. Kochenderfer, Tim A. Wheeler, March 12, 2019

[27] "Improving the Rprop Learning Algorithm",by Christian Igel , Michael Hüsken, PROCEEDINGS OF THE SECOND INTERNATIONAL SYMPOSIUM ON NEURAL COMPUTATION (NC 2000)

[28] "Adam: A Method for Stochastic Optimization" ,Diederik P. Kingma, Jimmy Ba,

Published as a conference paper at ICLR 2015

[29] "ON THE CONVERGENCE OF ADAM AND BEYOND",Sashank J. Reddi, Satyen Kale & Sanjiv Kumar, Published as a conference paper at ICLR 2018

[30] "Survey on categorical data for neural networks" ,John T. Hancock & Taghi M. Khoshgoftaar Journal of Big Data volume 7, Article number: 28 (2020)