ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Ενισχυτική Μάθηση - Ms Pacman

Reinforcement Learning – Ms Pacman

Από

Ζωή Ιωάννα Αθανασοπούλου

Υποβάλλεται
για την εκπλήρωση των προϋποθέσεων λήψης
Μεταπτυχιακού Διπλώματος
στην ειδίκευση «Προηγμένα Πληροφοριακά Συστήματα»
του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”
στο
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Ιούνιος 2024

Επιβλέπων/Επιβλέπουσα: Μιχαήλ Φιλιππάκης
Ακαδημαϊκή Θέση: Καθηγητής

Συγγραφέας Ζωή Ιωάννα Αθανασοπούλου

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή/Φοιτήτριας: Ζωή Ιωάννα Αθανασοπούλου

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας: Reinforcement Learning – Ms Pacman

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις 23/06/2024 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπων (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς)
Μιχαήλ Φιλιππάκης, Καθηγητής
Μέλος Εξεταστικής Επιτροπής: Ανδρέας Μενύχτας, Επίκουρος Καθηγητής
Μέλος Εξεταστικής Επιτροπής: Μαρία Χαλκίδη, Αναπληρώτρια Καθηγήτρια

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Η Ζωή Ιωάννα Αθανασοπούλου, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Reinforcement Learning – Agent-based games», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.
Επιπλέον *δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.*
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο/Η ΔΗΛΩΝ/ΟΥΣΑ
Ονοματεπώνυμο: Ζωή Ιωάννα Αθανασοπούλου
Αριθμός Μητρώου: ΜΕ2043
Υπογραφή:

Περιεχόμενα

Περίληψη....8
Abstract....9
1 Εισαγωγή....10
2 Ενισχυτική Μάθηση....11
2.1 Βασικά Στοιχεία....11
2.2 Το πρόβλημα της Ενισχυτικής Μάθησης....14
2.2.1 Διεπαφή Πράκτορα-Περιβάλλοντος....14
2.2.2 Το δίλημμα μεταξύ εξερεύνησης ή αξιοποίησης....15
2.2.3 Στόχοι και Ανταμοιβές....17
2.2.4 Η έννοια της Επιστροφής....18
2.2.5 Διαδικασίες Markov....20
2.2.6 Συναρτήσεις Αξίας....22
2.3 Μέθοδοι Ενισχυτικής Μάθησης....26
2.3.1 Δυναμικός Προγραμματισμός....26
2.3.2 Monte Carlo....27
2.3.3 Μάθηση Χρονικών Διαφορών....27
2.3.4 Ίχνη Επιλεξιμότητας....28
2.3.5 Μέθοδοι βασισμένοι σε μοντέλο....28
2.3.6 Ενοποιημένη Άποψη των μεθόδων....30
2.4 Γενίκευση-Προσέγγιση Συναρτήσεων....31
2.5 Ιεραρχική Ενισχυτική Μάθηση....31
2.5.1 Ημι-Μαρκοβιανές διαδικασίες απόφασης....32
2.5.2 Προσεγγίσεις....33
2.6 Σχεσιακή Ενισχυτική Μάθηση....35
2.6.1 Στόχοι....36
2.6.2 Μέθοδοι....37
2.7 Ποιότητα της Διαδικασίας Μάθησης....37
3 Μηχανική Μάθηση και Ηλεκτρονικά Παιχνίδια....39
3.1 Εισαγωγή....39
3.2 AI και ηλεκτρονικά παιχνίδια....40
3.2.1 Η χρησιμότητα της AI....40
3.2.2 Κλασσικές Μέθοδοι AI....42
3.2.3 Ψευδομάθηση....43
3.3 Μηχανική Μάθηση και Αναλυτικά Παιχνίδια....44

3.4 Ζητήματα εφαρμογής σε εμπορικά παιχνίδια.....45
3.4.1 Πηγές Μάθησης.....45
3.4.2 Τρόπος Εκπαίδευσης.....47
3.5 Τεχνικές Μηχανικής Μάθησης.....48
3.5.1 Δένδρα Απόφασης.....48
3.5.2 Νευρωνικά Δίκτυα.....49
3.5.3 Γενετικοί Αλγόριθμοι.....49
3.5.4 Μάθηση κατά Bayes.....50
3.6 Ενισχυτική Μάθηση.....51
3.6.1 Γενικά.....51
3.6.2 Δυναμική Παραγωγή Σεναρίων.....51
3.7 Προοπτικές.....52
3.7.1 Γενικές.....52
3.7.2 Προοπτικές για τη Βιομηχανία.....52
3.7.3 Επιδράσεις στο Gameplay.....53
3.8 Άλλες Σχετικές Εφαρμογές.....54
4 Αλγόριθμος Ενισχυτικής Μάθησης σε περιβάλλον παιχνιδιού.....55
4.1 Python.....55
4.1.1 Βιβλιοθήκες.....56
4.2 Περιβάλλον.....59
5 Αποτελέσματα.....69
5.1 Πράκτορας DQN.....69
5.2 Πράκτορας DQN με batch normalization.....72
5.3 Πράκτορας duel DQN.....75
5.4 Πράκτορας Noisy DQN.....77
5.5 Πράκτορας double DQN με prioritized experience buffer.....79
6 Συζήτηση.....82
6.1 Απόδοση των Πρακτόρων.....82
6.2 Ταχύτητα Εκπαίδευσης των Πρακτόρων.....83
6.3 Συγκριτική Ανάλυση.....83
6.4 Συμπεράσματα για την Ταχύτητα Εκπαίδευσης και την Απόδοση.....84
7 Συμπεράσματα.....85
7.1 Ανακεφαλαίωση Αποτελεσμάτων.....85
7.2 Θεωρητική Ανάλυση.....86
7.3 Συμπεράσματα για την Ταχύτητα Εκπαίδευσης και την Απόδοση.....86

7.4 Ανακεφαλαίωση.....86
Βιβλιογραφία.....87

Κατάλογος Εικόνων

Εικόνα 1: Ενοποιημένη επισκόπηση των μεθόδων ενισχυτικής μάθησης (Πηγή: Lanctot et al., 2017). 30
Εικόνα 2: Μάθηση μέσω παρατήρησης ανθρώπινης συμπεριφοράς (Πηγή: Bertolini et al., 2021)....46
Εικόνα 3: Μάθηση μέσω καθοδήγησης (Πηγή: Bertolini et al., 2021).....46
Εικόνα 4: Μάθηση μέσω εμπειρίας (Πηγή: Bertolini et al., 2021).....47
Εικόνα 5 Ένα καρέ του παιχνιδιού MsPacman.....60
Εικόνα 6. Δύο διαδοχικές καταστάσεις του περιβάλλοντος του παιχνιδιού, όπου τα εικονοστοιχεία του κόκκινου και του μοβ φαντάσματος τρεμοπαίζουν.....63
Εικόνα 7. Ένα καρέ του παιχνιδιού σε Grayscale.....64
Εικόνα 8. Περικομμένο καρέ.....65
Εικόνα 9: Η αρχιτεκτονική του δικτύου DQN.....66
Εικόνα 10. Το δίκτυο Duel DQN.....67
Εικόνα 11 Η μέση ανταμοιβή για κάθε δέκα επεισόδια για τον double DQN.....70
Εικόνα 12 Το σκορ του πράκτορα DQN.....70
Εικόνα 13 Η ταχύτητα εκπαίδευσής του πράκτορα DQN.....71
Εικόνα 14 Σύγκριση του εκπαιδευμένου πράκτορα DQN με έναν ανεκπαίδευτο.....72
Εικόνα 15 Η μέση ανταμοιβή για κάθε δέκα επεισόδια για τον πράκτορα double DQN με batch normalization.....72
Εικόνα 16 Το σκορ του πράκτορα DQN με batch normalization.....73
Εικόνα 17 Η ταχύτητα εκπαίδευσής του πράκτορα DQN με batch normalization.....74
Εικόνα 18Σύγκριση του εκπαιδευμένου πράκτορα DQN με batch normalization με έναν ανεκπαίδευτο.....74
Εικόνα 19 Η μέση ανταμοιβή για κάθε δύο επεισόδια για τον πράκτορα Duel DQN.....75
Εικόνα 20 Το σκορ του πράκτορα Duel DQN.....76
Εικόνα 21 Η ταχύτητα εκπαίδευσής του πράκτορα Duel DQN.....76
Εικόνα 22 Σύγκριση του εκπαιδευμένου πράκτορα duel DQN με έναν ανεκπαίδευτο.....77
Εικόνα 23 Η μέση ανταμοιβή για κάθε δύο επεισόδια για τον πράκτορα Noisy DQN.....78
Εικόνα 24 Το σκορ του πράκτορα Noisy DQN.....78
Εικόνα 25 Η ταχύτητα εκπαίδευσής του πράκτορα Noisy DQN.....79
Εικόνα 26 Σύγκριση του εκπαιδευμένου πράκτορα noisy DQN με έναν ανεκπαίδευτο.....79
Εικόνα 27 Η μέση ανταμοιβή για κάθε δύο επεισόδια για τον πράκτορα double DQN με prioritized experience buffer.....80
Εικόνα 28. Το σκορ του πράκτορα DQN με prioritized experience buffer.....80
Εικόνα 29 Η ταχύτητα εκπαίδευσής του πράκτορα DQN με prioritized replay buffer.....81
Εικόνα 30 Σύγκριση του εκπαιδευμένου πράκτορα DQN με prioritized replay buffer με έναν ανεκπαίδευτο.....81

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει την εφαρμογή της ενισχυτικής μάθησης για την εκπαίδευση πρακτόρων σε ηλεκτρονικά παιχνίδια. Σκοπός της έρευνας είναι να διερευνηθεί η απόδοση και η αποτελεσματικότητα διάφορων παραλλαγών του αλγορίθμου Deep Q-Network (DQN) σε περιβάλλοντα παιχνιδιών arcade.

Στην εργασία αυτή, αρχικά παρουσιάζονται οι βασικές έννοιες της ενισχυτικής μάθησης, συμπεριλαμβανομένων των πολιτικών, των συναρτήσεων ανταμοιβής και αξίας, και των μοντέλων περιβάλλοντος. Στη συνέχεια, αναλύονται οι θεωρητικές αρχές των αλγορίθμων DQN και οι επεκτάσεις τους, όπως το DQN με batch normalization, το Duel DQN, το Noisy DQN, και το Double DQN με Prioritized Experience Buffer.

Τα πειραματικά αποτελέσματα δείχνουν ότι κάθε παραλλαγή προσφέρει συγκεκριμένα πλεονεκτήματα στην εκπαίδευση των πρακτόρων. Οι πράκτορες που εκπαιδεύτηκαν με τις παραλλαγές του DQN παρουσίασαν βελτίωση στη μέση ανταμοιβή, στο σκορ και στην ταχύτητα εκπαίδευσης σε σχέση με τον βασικό αλγόριθμο DQN. Ειδικότερα, ο πράκτορας με Prioritized Experience Buffer και Double DQN εμφάνισε την καλύτερη συνολική απόδοση, επιτυγχάνοντας την υψηλότερη μέση ανταμοιβή και σκορ.

Η εργασία καταλήγει στο συμπέρασμα ότι η ενισχυτική μάθηση και οι προσαρμογές των αλγορίθμων DQN μπορούν να εφαρμοστούν αποτελεσματικά για την εκπαίδευση πρακτόρων σε δυναμικά και σύνθετα περιβάλλοντα, όπως τα ηλεκτρονικά παιχνίδια. Η θεωρητική ανάλυση και τα πειραματικά αποτελέσματα υποδεικνύουν τη σημαντική συμβολή των τεχνικών βελτιστοποίησης στη βελτίωση της αποδοτικότητας και της απόδοσης των πρακτόρων.

Λέξεις-κλειδιά: Ενισχυτική Μάθηση, Deep Q-Network, Batch Normalization, Duel DQN, Noisy DQN, Prioritized Experience Buffer, Ηλεκτρονικά Παιχνίδια, Εκπαίδευση, *Google Colab, GPU, MsPacman, Python, PyTorch, Παιχνίδι Arcade, Πράκτορας, Azure*

Abstract

This thesis explores the application of reinforcement learning for training agents in video games. The purpose of the research is to investigate the performance and effectiveness of various Deep Q-Network (DQN) algorithm variations in arcade game environments.

Initially, the thesis presents the basic concepts of reinforcement learning, including policies, reward functions, value functions, and environment models. Then, it analyzes the theoretical principles of DQN algorithms and their extensions, such as DQN with batch normalization, Duel DQN, Noisy DQN, and Double DQN with Prioritized Experience Buffer.

Experimental results show that each variation offers specific advantages in agent training. Agents trained with the DQN variations showed improvements in average reward, score, and training speed compared to the basic DQN algorithm. Specifically, the agent with Prioritized Experience Buffer and Double DQN demonstrated the best overall performance, achieving the highest average reward and score.

The thesis concludes that reinforcement learning and the adaptations of DQN algorithms can be effectively applied to train agents in dynamic and complex environments, such as video games. The theoretical analysis and experimental results highlight the significant contribution of optimization techniques in improving the efficiency and performance of the agents.

Keywords: Reinforcement Learning, Deep Q-Network, Batch Normalization, Duel DQN, Noisy DQN, Prioritized Experience Buffer, Video Games, Training, *Google Colab, GPU, MsPacman, Python, PyTorch, Arcade Game, Agent, Azure*

1 Εισαγωγή

Ο σκοπός αυτής της εργασίας είναι να αναδείξει την χρησιμότητα των ηλεκτρονικών παιχνιδιών ως πεδία έρευνας για την ενισχυτική μάθηση και την τεχνητή νοημοσύνη. Η ικανότητα μάθησης αποτελεί βασικό χαρακτηριστικό κάθε οντότητας που επιδιώκει να είναι έξυπνη. Συνήθως, οι ερευνητές στον τομέα της τεχνητής νοημοσύνης χρησιμοποιούν συνηθισμένα προβλήματα για να αποδείξουν την αποτελεσματικότητα των επιτευγμάτων τους, αν και αυτό μπορεί να διευκολύνει την ανάλυση και την εξαγωγή συμπερασμάτων.

Η ενισχυτική μάθηση αποτελεί μία από τις πιο συναρπαστικές πτυχές της τεχνητής νοημοσύνης και της μηχανικής μάθησης σήμερα. Από την προσομοίωση της διαδικασίας εκμάθησης των ανθρώπων έως την αυτόνομη λήψη αποφάσεων από πράκτορες σε πολύπλοκα περιβάλλοντα, η ενισχυτική μάθηση έχει αλλάξει τον τρόπο που οι υπολογιστές αποκτούν γνώση και λαμβάνουν αποφάσεις. Ένας σημαντικός τομέας εφαρμογής της ενισχυτικής μάθησης είναι η ανάπτυξη προγραμμάτων που μπορούν να εκπαιδευτούν να παίζουν παιχνίδια, όπως τα arcade games.

Σε αυτήν την εργασία επικεντρωνόμαστε στη χρήση ενισχυτικής μάθησης για την εκπαίδευση ενός πράκτορα ώστε να είναι επιδέξιος σε ένα arcade game. Χρησιμοποιώντας τη γλώσσα προγραμματισμού Python και τεχνικές ενισχυτικής μάθησης, στοχεύουμε στη δημιουργία ενός προγράμματος που μπορεί να μάθει και να βελτιώσει τις ικανότητές του στο παιχνίδι. Αυτή η προσέγγιση μας επιτρέπει να εξερευνήσουμε τις δυνατότητες και τα όρια της ενισχυτικής μάθησης.

Συνολικά, παρουσιάζουμε μια ολοκληρωμένη διαδικασία εκπαίδευσης ενός πράκτορα μέσω ενισχυτικής μάθησης, εστιάζοντας στην υλοποίηση και αξιολόγηση του προγράμματός μας σε ένα περιβάλλον arcade game σε Python. Μέσω αυτής της προσπάθειας, επιδιώκουμε να κατανοήσουμε τις δυνατότητες και τα όρια της ενισχυτικής μάθησης.

2 Ενισχυτική Μάθηση

Η Ενισχυτική Μάθηση ανήκει σε ένα από τα πιο δημοφιλή και ενδιαφέροντα πεδία της τεχνητής νοημοσύνης και της μηχανικής μάθησης. Το ενδιαφέρον αυτό έχει αυξηθεί δραματικά τα τελευταία χρόνια στην επιστημονική κοινότητα. Η ενισχυτική μάθηση διαμορφώθηκε από ένα ευρύ φάσμα επιστημονικών πεδίων, συμπεριλαμβανομένης της κυβερνητικής, της στατιστικής, της ψυχολογίας, της νευροβιολογίας και της επιστήμης των υπολογιστών. Αυτή η ποικιλομορφία επιστημονικών πηγών έχει καταστήσει την ενισχυτική μάθηση μια κοινή γλώσσα, επιτρέποντας σε επιστήμονες από διάφορα πεδία να μοιράζονται τα προβλήματα και τις ανακαλύψεις τους (Yang & Wang, 2020).

Η ενισχυτική μάθηση ουσιαστικά παρέχει ένα υπολογιστικό πλαίσιο για την εκμάθηση συμπεριφορών από ευφυείς πράκτορες, βασιζόμενο σε ένα σήμα ανταμοιβής. Αυτή η ανταμοιβή μπορεί να είναι οτιδήποτε από τροφή, νερό, χρήματα ή οποιοδήποτε άλλο μέτρο επίδοσης του πράκτορα στην εκμάθηση των επιθυμητών συμπεριφορών (Li, 2017).

Ένας από τους κύριους στόχους των ερευνητών στο πεδίο της ενισχυτικής μάθησης είναι να αναζητήσουν τρόπους για να προγραμματίσουν ευφυείς πράκτορες, παρέχοντας τους απλώς ανταμοιβές και τιμωρίες, χωρίς να πρέπει να προσδιορίσουν ακριβώς πώς θα επιτύχουν τις εργασίες που τους αναθέτουν. Παρόλο που η επίτευξη αυτού του στόχου απαιτεί μεγάλη υπολογιστική πολυπλοκότητα, υπάρχουν επιτυχημένα παραδείγματα, όπως ο πράκτορας που δημιούργησε ο Tesauro, ο οποίος κατάφερε να ανταγωνιστεί επιτυχώς παγκόσμιους πρωταθλητές στο παιχνίδι τάβλι.

Η Ενισχυτική Μάθηση αναφέρεται στην προσπάθεια ενός ευφυούς πράκτορα να μάθει μια συμπεριφορά μέσω δοκιμής και αποτυχίας σε ένα δυναμικό περιβάλλον. Μια προσέγγιση του προβλήματος είναι η αναζήτηση στο χώρο των πιθανών συμπεριφορών για να εντοπιστεί μια ενέργεια που θα οδηγήσει σε επιθυμητά αποτελέσματα σε μια συγκεκριμένη κατάσταση του περιβάλλοντος. Αυτή η προσέγγιση μοιάζει με αυτήν που ακολουθείται στον γενετικό προγραμματισμό και τους γενετικούς αλγόριθμους. Ωστόσο, η κυρίαρχη προσέγγιση στην Ενισχυτική Μάθηση είναι η χρήση μεθόδων δυναμικού προγραμματισμού και στατιστικής για να αξιολογηθεί η επίδοση μιας ενέργειας ανά χρονική στιγμή (Yang & Wang, 2020).

2.1 Βασικά Στοιχεία

Στην Ενισχυτική Μάθηση, το βασικό θέμα εστίασης είναι ένας πράκτορας που αλληλεπιδρά με ένα περιβάλλον. Ο πράκτορας λαμβάνει ενέργειες που τροποποιούν την κατάσταση του περιβάλλοντος,

το οποίο με τη σειρά του φέρνει νέες καταστάσεις στον πράκτορα (Levine et al., 2020).

Εκτός από τον πράκτορα και το περιβάλλον, υπάρχουν 4 βασικές έννοιες στην ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ:

- Πολιτική (policy): Ο κανόνας που καθορίζει ποια ενέργεια θα επιλέξει ο πράκτορας σε κάθε κατάσταση.

- Συνάρτηση ανταμοιβής (reward function): Ορίζει την ανταμοιβή που λαμβάνει ο πράκτορας για κάθε μετάβαση σε νέα κατάσταση.

- Συνάρτηση αξίας (value function): Εκτιμά πόσο ωφέλιμη είναι η κάθε κατάσταση για τον πράκτορα, λαμβάνοντας υπόψη μελλοντικές ανταμοιβές.

- Μοντέλο περιβάλλοντος (environment model): (Προαιρετικό) Μια προσομοίωση του περιβάλλοντος που βοηθά τον πράκτορα να προβλέψει τις συνέπειες των ενεργειών του.

Κατανοώντας αυτές τις βασικές έννοιες, θέτονται τα θεμέλια για την υλοποίηση αλγορίθμων Ενισχυτικής Μάθησης και την επίλυση σύνθετων προβλημάτων (Levine et al., 2020).

Η πολιτική στην Ενισχυτική Μάθηση καθορίζει τον τρόπο συμπεριφοράς του πράκτορα σε κάθε συγκεκριμένη χρονική στιγμή. Από πλευράς εφαρμογής, η πολιτική είναι συχνά μια αντιστοίχιση μεταξύ των καταστάσεων ή των παρατηρήσεων που αντιλαμβάνεται ο πράκτορας και των ενεργειών που επιλέγει σε αυτές τις καταστάσεις. Σε ορισμένες περιπτώσεις, μια πολιτική μπορεί να απλοποιηθεί σε έναν πίνακα αντιστοίχισης, γνωστό και ως πίνακας αναζήτησης (lookup table), ενώ σε άλλες περιπτώσεις ο ορισμός της μπορεί να απαιτεί πιο περίπλοκες υπολογιστικές διαδικασίες. Αν και η πολιτική αποτελεί τον κύριο πυρήνα ενός πράκτορα στην Ενισχυτική Μάθηση και είναι επαρκής για την καθοριστική επίδραση της συμπεριφοράς του, συνήθως είναι στοχαστική, δηλαδή καθορίζει τις πιθανότητες επιλογής των ενεργειών από τον πράκτορα σε κάθε κατάσταση. Από την άλλη πλευρά, η συνάρτηση ανταμοιβής καθορίζει τον στόχο σε ένα πρόβλημα Ενισχυτικής Μάθησης.

Η συνάρτηση ανταμοιβής στο πλαίσιο της Ενισχυτικής Μάθησης ορίζει τον στόχο του προβλήματος. Κάθε κατάσταση του περιβάλλοντος ή ζεύγος κατάστασης-ενέργειας αντιστοιχίζεται σε έναν αριθμό, την ανταμοιβή, η οποία δείχνει πόσο επιθυμητό είναι να βρίσκεται ο πράκτορας σε αυτή την κατάσταση ή να επιλέγει τη συγκεκριμένη ενέργεια, αντίστοιχα. Η κύρια αποστολή ενός πράκτορα στην Ενισχυτική Μάθηση είναι να μεγιστοποιήσει τη συνολική ανταμοιβή που λαμβάνει μακροπρόθεσμα. Ο πράκτορας μπορεί να χρησιμοποιήσει τη συνάρτηση ανταμοιβής για να προσαρμόσει την πολιτική του. Για παράδειγμα, αν μια ενέργεια που επιλέγεται με βάση την τρέχουσα πολιτική παράγει χαμηλή ανταμοιβή, τότε η πολιτική μπορεί να τροποποιηθεί ώστε να επιλέγει μια άλλη

ενέργεια σε αντίστοιχες περιστάσεις στο μέλλον. Γενικά, όπως και οι πολιτικές, οι συναρτήσεις ανταμοιβής είναι στοχαστικές.

Αντίθετα με τη συνάρτηση ανταμοιβής που καθορίζει τι είναι επιθυμητό αμέσως, μια συνάρτηση αξίας προσδιορίζει τι είναι επιθυμητό μακροπρόθεσμα. Η αξία μιας κατάστασης ορίζεται ως το συνολικό ποσό ανταμοιβής που μπορεί να συγκεντρώσει μελλοντικά ο πράκτορας, ξεκινώντας από αυτή την κατάσταση. Η αξία μιας κατάστασης δείχνει πόσο επιθυμητή είναι αυτή η κατάσταση μακροπρόθεσμα, λαμβάνοντας υπόψιν τις καταστάσεις που θα ακολουθήσουν και τις διαθέσιμες ανταμοιβές που μπορούν να προκύψουν κατά τη μετάβαση προς αυτές τις καταστάσεις και από αυτές. Για παράδειγμα, μπορεί μια κατάσταση να παρέχει πάντα χαμηλή ανταμοιβή, αλλά η αξία της μπορεί να είναι υψηλή επειδή συνήθως ακολουθείται από καταστάσεις που προσφέρουν υψηλές ανταμοιβές (ή ακόμα και το αντίστροφο) (Levine et al., 2020).

Οι ανταμοιβές θεωρούνται ουσιώδεις, ενώ οι αξίες, που αναγνωρίζουν και προβλέπουν τις ανταμοιβές, έχουν δευτερεύουσα σημασία. Χωρίς τις ανταμοιβές, οι αξίες δεν θα μπορούσαν να υπάρξουν, καθώς η μοναδική αιτία για την εκτίμηση των αξιών είναι η προσπάθεια να αποκτηθεί περισσότερη ανταμοιβή. Ωστόσο, κατά τη λήψη και την αξιολόγηση αποφάσεων, γίνεται εστίαση στις αξίες. Οι αποφάσεις για τις ενέργειες βασίζονται σε εκτιμήσεις των αξιών, με στόχο την αναζήτηση ενεργειών που οδηγούν σε καταστάσεις με τη μέγιστη δυνατή αξία, καθώς αυτές οι ενέργειες παρέχουν τη μέγιστη ανταμοιβή μακροπρόθεσμα. Ωστόσο, η εκτίμηση των αξιών είναι πολύ πιο δύσκολη από την εκτίμηση των ανταμοιβών, οι οποίες προέρχονται άμεσα από το περιβάλλον. Οι αξίες πρέπει να εκτιμηθούν και να επανεκτιμηθούν μέσω των παρατηρήσεων που κάνει ένας πράκτορας με την πάροδο του χρόνου. Ουσιαστικά, η αποτελεσματική εκτίμηση των αξιών είναι το κυριότερο στοιχείο και στόχος σχεδόν όλων των αλγορίθμων Ενισχυτικής Μάθησης.

Το μοντέλο του περιβάλλοντος αποτελεί μια επικεντρωμένη τεχνητή οντότητα που επιδιώκει να αντιγράψει τη συμπεριφορά του περιβάλλοντος. Για παράδειγμα, μπορεί να χρησιμοποιηθεί για την πρόβλεψη της ανταμοιβής με βάση μια κατάσταση και μια ενέργεια. Τα μοντέλα αυτά χρησιμοποιούνται για το σχεδιασμό ενεργειών, λαμβάνοντας υπόψη τις μελλοντικές καταστάσεις πριν ακόμα αντιμετωπιστούν από τον πράκτορα. Αρχικά, η Έξυπνη Μηχανή θεωρήθηκε αντίθετη με το σχεδιασμό ενεργειών, αλλά στη συνέχεια παρουσιάστηκαν προσεγγίσεις που συνδύαζαν τη μάθηση μέσω δοκιμής και αποτυχίας, τη μάθηση ενός μοντέλου για το περιβάλλον, και τη χρήση αυτού για τον σχεδιασμό ενεργειών. Αυτή η σύγκλιση κυρίως οφείλεται στην αποσαφήνιση της σχέσης μεταξύ της Έξυπνης Μηχανής και του δυναμικού προγραμματισμού που χρησιμοποιεί μοντέλα, τα οποία, αντίστοιχα, σχετίζονται στενά με μεθόδους σχεδιασμού ενεργειών σε χώρους καταστάσεων. Οι σύγχρονες μέθοδοι της Έξυπνης Μηχανής καλύπτουν ένα ευρύ φάσμα, από απλή μάθηση μέσω δοκι-

μής και αποτυχίας (χωρίς τη χρήση μοντέλων) έως υψηλού επιπέδου εκούσιο σχεδιασμό ενεργειών (Levine et al., 2020).

2.2 Το πρόβλημα της Ενισχυτικής Μάθησης

Σε αυτή την ενότητα, περιγράφονται τα βασικά στοιχεία που σχετίζονται με την Ενισχυτική Μάθηση. Αρχικά, παρουσιάζεται το γενικό πλαίσιο της ενισχυτικής μάθησης. Στη συνέχεια, αναφέρεται το δίλημμα μεταξύ εξερεύνησης και εκμετάλλευσης των πληροφοριών. Κατόπιν, εξετάζονται οι έννοιες της ανταμοιβής και της επιστροφής. Τέλος, παρουσιάζονται οι Μαρκοβιανές Διαδικασίες Λήψης Αποφάσεων και η κρίσιμης σημασίας συναρτήσεις αξίας.

2.2.1 Διεπαφή Πράκτορα-Περιβάλλοντος

Ο ορισμός του προβλήματος στην Ενισχυτική Μάθηση διαμορφώνεται με τρόπο που παρέχει ένα άμεσο πλαίσιο για την εκπαίδευση μέσω δοκιμής και αποτυχίας προκειμένου να επιτευχθεί ένας στόχος. Ο υποκείμενος που μαθαίνει και λαμβάνει αποφάσεις αναφέρεται ως πράκτορας, ενώ το περιβάλλον αποτελείται από όλες τις άλλες οντότητες εκτός του πράκτορα. Ο πράκτορας αλληλεπιδρά με το περιβάλλον, επιλέγοντας ενέργειες και το περιβάλλον ανταποκρίνεται παρέχοντας ανταμοιβές. Αυτή η αλληλεπίδραση οδηγεί σε νέες καταστάσεις, με το περιβάλλον να παρέχει αριθμητικές ανταμοιβές που ο πράκτορας επιθυμεί να μεγιστοποιήσει στο μακροπρόθεσμο. Κάθε στιγμή της αλληλεπίδρασης αυτής, ο πράκτορας αντιλαμβάνεται την τρέχουσα κατάσταση του περιβάλλοντος, επιλέγει μια ενέργεια βάσει αυτής της κατάστασης, λαμβάνει μια ανταμοιβή και μεταβαίνει σε μια νέα κατάσταση (Szita, 2012).

Κάθε φορά, σε κάθε χρονικό βήμα, ο πράκτορας διαμορφώνει μια αντιστοίχιση μεταξύ των καταστάσεων του περιβάλλοντος και των πιθανοτήτων επιλογής των διαθέσιμων ενεργειών. Αυτή η αντιστοίχιση ονομάζεται πολιτική και συμβολίζεται με π, όπου π(s, a) είναι η πιθανότητα να επιλεγεί η ενέργεια α όταν η κατάσταση είναι s, σε ένα συγκεκριμένο χρονικό σημείο t. Οι μέθοδοι Ενισχυτικής Μάθησης καθορίζουν πώς οι πράκτορες προσαρμόζουν την πολιτική τους με βάση τις εμπειρίες που αποκτούν (Dulac-Arnold, Mankowitz & Hester, 2019).

Αυτό το πλαίσιο είναι γενικό και ευέλικτο, επιτρέποντας τη χρήση του σε πολλά διαφορετικά προβλήματα και με ποικίλους τρόπους. Για παράδειγμα, τα χρονικά βήματα δεν απαιτείται να αντιστοιχούν σε σταθερά χρονικά διαστήματα, αλλά μπορεί να αναφέρονται σε διαδοχικές φάσεις της δια-

δικασίας λήψης αποφάσεων. Επίσης, οι καταστάσεις μπορεί να ποικίλλουν από πρωταρχικές πληροφορίες μέχρι υψηλού επιπέδου αφαιρετικές πληροφορίες. Η αντίληψη του πράκτορα για το περιβάλλον μπορεί να είναι περιορισμένη, οδηγώντας σε μερική παρατηρησιμότητα του περιβάλλοντος. Όσον αφορά τις ενέργειες, αυτές μπορούν να κυμαίνονται από βασικές ενέργειες χαμηλού επιπέδου έως μακροενέργειες που αποτελούνται από ακολουθίες ενεργειών χαμηλού επιπέδου, καθώς και αποφάσεις υψηλού επιπέδου.

Επιπλέον, η διαχωριστική γραμμή μεταξύ του περιβάλλοντος και του πράκτορα συνήθως δεν συμπίπτει με το φυσικό σύνορο του φυσικού σώματός τους, όπως συμβαίνει για παράδειγμα με ζώα ή ρομπότ. Συνήθως θεωρείται ότι αυτό το όριο βρίσκεται πιο κοντά στον πράκτορα απ' ό,τι σε σχέση με τη νοοτροπία που διέπει την παραπάνω θεώρηση. Για παράδειγμα, οι κινητήρες, οι μηχανικοί σύνδεσμοι και τα αισθητήρια μηχανήματα ενός ρομπότ συνήθως θεωρούνται μέρος του περιβάλλοντος του πράκτορα παρά συστατικά του. Επιπλέον, οι ανταμοιβές υπολογίζονται εντός του φυσικού σώματος του πράκτορα, αλλά θεωρούνται εξωτερικές ως προς αυτόν. Ο γενικός κανόνας που ακολουθείται είναι ότι οτιδήποτε δεν μπορεί να επηρεαστεί ή να αλλαχθεί αυθαίρετα από τον πράκτορα θεωρείται εξωτερικό ως προς αυτόν και, συνεπώς, μέρος του περιβάλλοντός του. Το πλαίσιο της Ενισχυτικής Μάθησης μπορεί να θεωρηθεί ως μια αξιοσημείωτη αφαιρετική αναπαράσταση του προβλήματος της προσανατολισμένης σε στόχους μάθησης, μέσω αλληλεπίδρασης. Σύμφωνα με αυτό, το πρόβλημα μπορεί να αναχθεί σε τρία σήματα που διασχίζουν τον πράκτορα και το περιβάλλον (Dulac-Arnold, Mankowitz & Hester, 2019):

- ένα σήμα που αναπαριστά τις επιλογές του πράκτορα (ενέργειες)
- ένα σήμα που αναπαριστά τη βάση για τις επιλογές αυτές (καταστάσεις) και
- ένα σήμα που καθορίζει τον στόχο του πράκτορα (ανταμοιβές).

Αν και αυτό το πλαίσιο ενδέχεται να μην είναι επαρκές για την αναπαράσταση όλων των προβλημάτων μάθησης απόφασης, έχει αποδειχθεί ευρέως χρήσιμο και εφαρμόσιμο στην πράξη.

2.2.2 Το δίλημμα μεταξύ εξερεύνησης ή αξιοποίησης

Μια βασική διαφορά της Ενισχυτικής Μάθησης σε σχέση με την επιβλεπόμενη μάθηση είναι η ανάγκη του πράκτορα να εξερευνήσει το περιβάλλον του. Αυτό προκαλεί ένα δίλημμα μεταξύ εξερεύνησης και αξιοποίησης. Ο πράκτορας μπορεί να πιστεύει ότι η επιλογή μιας ενέργειας θα οδηγήσει σε υψηλή ανταμοιβή, αλλά πρέπει να αποφασίσει αν πρέπει να την επιλέξει κάθε φορά ή αν θα

ήταν καλύτερο να εξετάσει και άλλες επιλογές, παρά την έλλειψη πληροφοριών γι' αυτές. Η απάντηση σε αυτό το ερώτημα εξαρτάται από την περίοδο αλληλεπίδρασης του πράκτορα με το περιβάλλον. Όσο περισσότερο διαρκεί η αλληλεπίδραση, τόσο πιο αρνητικές είναι οι συνέπειες της πρόωρης σύγκλισης σε μια μη βέλτιστη συμπεριφορά και, συνεπώς, τόσο πιο σημαντική είναι η ανάγκη για εξερεύνηση. Συνήθως, για να αντιμετωπιστεί αυτό το δίλημμα, χρησιμοποιούνται απλοί αλγόριθμοι εξερεύνησης. Ένα σχετικό ευριστικό που μπορεί να βοηθήσει είναι η αισιοδοξία σε περίπτωση αβεβαιότητας. Σύμφωνα με αυτήν, ο πράκτορας επιλέγει ενέργειες με βάση αισιόδοξες προηγούμενες πεποιθήσεις για τις ανταμοιβές τους, ενώ απαιτείται η ύπαρξη ισχυρών αρνητικών στοιχείων για να αλλάξει αυτή η επιλογή (Zhang, Li, Shi & Hwang, 2022).

Μια από τις πιο δημοφιλείς και απλές μεθόδους είναι η ε-γκάρφεια (ε-greedy), η οποία αποτελεί μια παραλλαγή της άπληστης επιλογής ενεργειών που αναφέρθηκε προηγουμένως. Σύμφωνα με αυτήν τη μέθοδο, ο πράκτορας επιλέγει την ενέργεια που θεωρείται βέλτιστη βάσει της τρέχουσας πολιτικής με μια πιθανότητα 1-ε και μια τυχαία ενέργεια με πιθανότητα ε. Σε ορισμένες περιπτώσεις, αυτή η μέθοδος υλοποιείται με τον καθορισμό μιας αρκετά υψηλής τιμής ε, η οποία στη συνέχεια μειώνεται με αργό ρυθμό, προκειμένου να προωθηθεί η εξερεύνηση κυρίως στα αρχικά στάδια της διαδικασίας μάθησης. Ένα σημαντικό μειονέκτημα αυτής της μεθόδου είναι ότι όταν επιλέγει μια μη-βέλτιστη ενέργεια, το κάνει εντελώς τυχαία, χωρίς να μπορεί να διακρίνει μια υποσχόμενη εναλλακτική από μια που έχει αποδειχτεί ότι είναι ανεπιθύμητη.

Τέλος, μια άλλη δημοφιλής τεχνική για την αντιμετώπιση του διλήμματος μεταξύ εξερεύνησης και αξιοποίησης είναι η εξερεύνηση Softmax. Αυτή η τεχνική στοχεύει στην επίλυση του προβλήματος που παρουσιάζει η ε-γκάρφεια, καθορίζοντας τις πιθανότητες επιλογής κάθε ενέργειας βάσει μιας κλιμακούμενης συνάρτησης της αναμενόμενης αξίας. Σε αυτήν την προσέγγιση, η ενέργεια που εκτιμάται ως τρέχουσα βέλτιστη έχει τη μεγαλύτερη πιθανότητα επιλογής, αλλά όλες οι υπόλοιπες ενέργειες κατατάσσονται και τους αποδίδονται βάρη ανάλογα με τις εκτιμήσεις αξίας τους. Συνήθως, χρησιμοποιούνται κατανομές Gibbs ή Boltzmann, ενώ η επιλογή μιας ενέργειας α υπολογίζεται με την ακόλουθη πιθανότητα:

$$\frac{e^{Q_t(a)/\mathrm{T}}}{\sum_{b=1}^{n} e^{Q_t(b)/\mathrm{T}}} \quad (1)$$

όπου *T* είναι η παράμετρος που ονομάζεται θερμοκρασία. Υψηλές τιμές θερμοκρασίας οδηγούν σε ισοπίθανη επιλογή ενεργειών, ενώ χαμηλές τιμές οδηγούν σε μεγαλύτερη διαφορά στις πιθανότητες επιλογής μεταξύ ενεργειών που διαφέρουν στις εκ τιμήσεις αξίας τους. Στο όριο, όταν $T \rightarrow 0$, η μέθοδος εξερεύνησης Softmax μετατρέπεται στην άπληστη επιλογή ενεργειών (Zhang, Li, Shi &

Hwang, 2022).

2.2.3 Στόχοι και Ανταμοιβές

Όπως αναφέρθηκε προηγουμένως, ο σκοπός του πράκτορα ορίζεται με κάποιον τρόπο μέσω του σήματος ανταμοιβής που λαμβάνει από το περιβάλλον. Σε κάθε χρονικό βήμα t, η ανταμοιβή r_t είναι ένας αριθμός που ανήκει στο σύνολο των πραγματικών αριθμών ($r_t \in R$. Ουσιαστικά, ο στόχος του πράκτορα είναι να μεγιστοποιήσει το συνολικό άθροισμα των ανταμοιβών που λαμβάνει, προσανατολιζόμενος προς το μακροπρόθεσμο όφελος και όχι μόνο προς την άμεση ανταμοιβή σε κάθε βήμα (Florensa et al., 2018).

Η χρήση του σήματος ανταμοιβής για την καθορισμό των στόχων αποτελεί ένα από τα κυριότερα χαρακτηριστικά της Ενισχυτικής Μάθησης. Παρόλο που αυτή η μέθοδος μπορεί να φανεί περιοριστική, στην πραγματικότητα έχει αποδειχθεί ευέλικτη και εφαρμόσιμη. Παρακάτω παρατίθενται παραδείγματα που επιδεικνύουν την ισχύ αυτής της προσέγγισης:

- Σε ένα περιβάλλον όπου ένα ρομπότ μαθαίνει να περπατάει, υπάρχει η δυνατότητα ορισμού μιας ανταμοιβής σε κάθε βήμα που είναι ανάλογη της προώθησης του.

- Σε ένα περιβάλλον όπου το ρομπότ μαθαίνει να δραπετεύει από ένα λαβύρινθο, η ανταμοιβή μπορεί να οριστεί ως μηδέν μέχρις ότου το ρομπότ εξέλθει από το λαβύρινθο, οπότε και η ανταμοιβή γίνεται θετική.

- Εναλλακτικά, υπάρχει η δυνατότητα ορισμού αρνητικών ανταμοιβών για κάθε βήμα που το ρομπότ παραμένει εγκλωβισμένο στο λαβύρινθο, προκειμένου να ενθαρρυνθεί και να δραπετεύσει το συντομότερο δυνατόν.

Σε όλα αυτά τα παραδείγματα, ο στόχος του πράκτορα είναι πάντα η μεγιστοποίηση της ανταμοιβής του. Για να οριστούν οι στόχοι του πράκτορα, οι ανταμοιβές πρέπει να καθοριστούν με τέτοιον τρόπο ώστε ο πράκτορας, μέσω της μεγιστοποίησής τους, να επιτυγχάνει παράλληλα αυτούς τους στόχους. Συνεπώς, το σήμα της ανταμοιβής έχει σκοπό να καθορίσει στον πράκτορα ποιος είναι ο στόχος του, αλλά όχι τον τρόπο με τον οποίο θα τον επιτύχει. Αυτό σημαίνει ότι πρέπει να διαμορφώνεται με προσοχή, ώστε να αντανακλά ακριβώς τους επιθυμητούς στόχους του συστήματος. Αν οι ανταμοιβές δεν διαμορφωθούν σωστά, ο πράκτορας ενδέχεται να εστιάσει σε λανθασμένους στόχους που δεν συνδέονται άμεσα με το επιθυμητό αποτέλεσμα (Florensa et al., 2018).

Επιπλέον, η παράγωγη των ανταμοιβών πρέπει να λαμβάνει υπόψη την ευελιξία και τη δυνατότητα προσαρμογής τους σε διαφορετικά περιβάλλοντα και καταστάσεις. Αυτό είναι σημαντικό για την αποτελεσματικότητα του συστήματος σε ποικίλες συνθήκες και για την αποφυγή της υπερεκμηδένης συμπεριφοράς. Επιπλέον, η διαφοροποίηση των ανταμοιβών μπορεί να ενθαρρύνει την εξερεύνηση του χώρου λύσης και την ανάπτυξη ποικίλων στρατηγικών που ενδέχεται να οδηγήσουν σε βέλτιστες λύσεις.

Συνεπώς, η διαμόρφωση των ανταμοιβών είναι ένας κρίσιμος παράγοντας για την επιτυχή λειτουργία των αλγορίθμων Ενισχυτικής Μάθησης, καθώς επηρεάζει τη συμπεριφορά και την απόδοσή τους (Florensa et al., 2018).

2.2.4 Η έννοια της Επιστροφής

Όσον αφορά τον στόχο της μάθησης που έχει αναφερθεί έως τώρα, έχει επισημανθεί ότι αυτός συνίσταται στη μεγιστοποίηση του συνολικού ποσού της ανταμοιβής που λαμβάνει ο πράκτορας σε μακροπρόθεσμη βάση. Με βάση αυτήν την προσέγγιση, ο πράκτορας προσπαθεί να αυξήσει την αναμενόμενη επιστροφή, η οποία ορίζεται ως κάποια συνάρτηση των ανταμοιβών που λαμβάνει κατά την πορεία του. Στην πιο απλή μορφή της, αυτή η επιστροφή μπορεί να υπολογιστεί ως το άθροισμα των ανταμοιβών μέχρι το τελικό χρονικό σημείο.

Αυτή η προσέγγιση είναι περισσότερο κατάλληλη για εφαρμογές όπου υπάρχει έννοια τερματικού χρονικού σημείου. Σε αυτές τις περιπτώσεις, η αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος φυσικά διαχωρίζεται σε διακριτά επεισόδια, όπως παρτίδες ενός παιχνιδιού ή γύροι σε αγωνιστική πίστα (Almahdi & Yang, 2017).

Σε αντίθεση, σε πολλές περιπτώσεις η αλληλεπίδραση πράκτορα-περιβάλλοντος δεν διαχωρίζεται φυσικά σε διακριτά επεισόδια, αλλά συνεχίζεται ατελείωτα. Σε αυτές τις περιπτώσεις, η έννοια της έκπτωσης μπορεί να χρησιμοποιηθεί για να αντιμετωπιστεί το πρόβλημα της άπειρης επιστροφής. Με τη χρήση αυτής της προσέγγισης, η τρέχουσα αξία των μελλοντικών ανταμοιβών λαμβάνεται υπόψη, λαμβάνοντας υπόψη τη μείωση της αξίας τους με την πάροδο του χρόνου. Τέλος, για να αναπαρασταθεί η έννοια της ανταμοιβής σε επεισοδιακές ή συνεχιζόμενες διαδικασίες, μπορεί να χρησιμοποιηθεί μια ενιαία μαθηματική σημειολογία, όπου η αναμενόμενη επιστροφή υπολογίζεται με βάση το συνολικό άθροισμα των ανταμοιβών, με την έννοια της έκπτωσης να λαμβάνεται υπόψη όταν απαιτείται.

Αντίθετα, σε πολλές περιπτώσεις, η αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος δεν διαιρείται φυσικά σε διακριτά επεισόδια, αλλά συνεχίζεται ατελείωτα. Αυτές οι διαδικασίες ονομάζονται συνεχιζόμενες. Σε αυτήν την περίπτωση, η εξίσωση:

$$R_t = r_{t+1} + r_{t+2} + \ldots + r_T \tag{2}$$

γίνεται προβληματική, διότι στο τερματικό βήμα υπάρχει ως *T* το άπειρο, και η επιστροφή μπορεί να είναι άπειρη και, συνεπώς, απροσδιόριστη. Για να αντιμετωπιστεί αυτό το πρόβλημα, εισάγεται η έννοια της έκπτωσης (Almahdi & Yang, 2017).

Σύμφωνα με αυτήν την προσέγγιση, ο πράκτορας προσπαθεί να επιλέγει ενέργειες έτσι ώστε το άθροισμα των μελλοντικών ανταμοιβών να μεγιστοποιείται μελλοντικά. Συγκεκριμένα, επιλέγει κάθε ενέργεια ώστε να μεγιστοποιείται η αναμενόμενη εκπτωθείσα επιστροφή:

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

όπου γ είναι μια παράμετρος με $0 \leq \gamma \leq 1$, που ονομάζεται ρυθμός έκπτωσης. Ο ρυθμός έκπτωσης επηρεάζει την τρέχουσα αξία των μελλοντικών ανταμοιβών, καθώς μια ανταμοιβή που λαμβάνεται σε μελλοντικά βήματα έχει αξία $\gamma^k - 1$ φορές μικρότερη από ό,τι θα είχε αν λαμβανόταν άμεσα. Εάν $\gamma \leq 1$, τότε το άπειρο άθροισμα έχει περιορισμένη τιμή, εφόσον η ακολουθία $\{r_k\}$ είναι φραγμένη. Εάν $\gamma = 0$, τότε ο πράκτορας είναι μυωπικός και ενδιαφέρεται μόνο για τη μεγιστοποίηση των άμεσων ανταμοιβών, ενώ αν το γ πλησιάζει στο 1, τότε ο πράκτορας λαμβάνει υπόψη του πιο σοβαρά και τις μελλοντικές ανταμοιβές (Almahdi & Yang, 2017).

Τέλος, μπορεί να χρησιμοποιηθεί μια ενιαία μαθηματική σημειολογία για να αναπαρασταθεί η έννοια της ανταμοιβής τόσο για επεισοδιακές όσο και για συνεχιζόμενες διαδικασίες. Σε αυτήν την περίπτωση, η ανταμοιβή ορίζεται ως:

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1} \tag{4}$$

όπου μπορεί να ισχύει $T = \infty$ ή $\gamma = 1$, αλλά ποτέ και τα δύο ταυτόχρονα (Almahdi & Yang, 2017).

2.2.5 Διαδικασίες Markov

Στο πλαίσιο της Ενισχυτικής Μάθησης, ο πράκτορας λαμβάνει τις αποφάσεις του βάσει του σήματος κατάστασης που του παρέχεται από το περιβάλλον. Είναι λογικό, συνεπώς, να υπάρχει η υπόθεση ότι η φύση αυτού του σήματος διαδραματίζει κρίσιμο ρόλο στη μαθησιακή διαδικασία. Ένα σήμα κατάστασης θα θεωρείται ιδανικό αν μπορούσε να παρέχει μια σύνοψη όλων των σχετικών πληροφοριών που αφορούν τα αισθητήρια σήματα, για όλες τις προηγούμενες χρονικές στιγμές. Για να επιτευχθεί αυτό, συνήθως απαιτούνται επιπλέον στοιχεία εκτός από αυτά που παρέχονται από την άμεση αντίληψη της κατάστασης από τον πράκτορα για κάποια χρονική στιγμή, αλλά σίγουρα όχι περισσότερα από το πλήρες ιστορικό των παρελθόντων αισθητήριων σημάτων κατάστασης. Ένα σήμα κατάστασης που μπορεί να διατηρεί όλες τις σχετικές πληροφορίες έχει την ιδιότητα της Markov (Markov Property) και ονομάζεται Markov. Για παράδειγμα, οι θέσεις των κομματιών σε μια παρτίδα σκάκι μπορούν να χρησιμοποιηθούν ως κατάσταση Markov, καθώς συνοψίζουν οτιδήποτε σημαντικό αφορά την ακολουθία κινήσεων των σκακιστών που έφερε την παρτίδα στην τρέχουσα κατάσταση (Wachi & Sui, 2020).

Στη συνέχεια, θα δοθεί μια τυπική οριοθέτηση της ιδιότητας Markov για το πρόβλημα της Ενισχυτικής Μάθησης, υποθέτοντας ότι το σύνολο των διαθέσιμων καταστάσεων και των τιμών των ανταμοιβών είναι πεπερασμένο. Γενικά, ένα περιβάλλον ανταποκρίνεται στη χρονική στιγμή *t + 1* στην ενέργεια που επιλέγει ο πράκτορας στη χρονική στιγμή *t*, και η απόκρισή του εξαρτάται από την πλήρη ακολουθία των παρελθόντων συμβάντων. Σε αυτήν την περίπτωση, οι δυναμικές του περιβάλλοντος μπορούν να καθοριστούν μόνο με τον πλήρη καθορισμό της κατανομής πιθανοτήτων, όπως περιγράφεται από την εξίσωση:

$$\mathbf{Pr} = \{s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t+1}, a_{t+1}, \ldots, r_1, s_0, a_0\} \quad (5)$$

για κάθε s_0, r και για όλες τις πιθανές τιμές των παρελθόντων γεγονότων, όπως περιγράφεται από την αλληλουχία s_t, a_t, r_t, s_{t+1}, a_{t+1}, ..., r_1, s_0, a_0. Εάν το σήμα κατάστασης είναι Markov, τότε η απόκριση του περιβάλλοντος εξαρτάται μόνο από τις αναπαραστάσεις της κατάστασης και της ενέργειας κατά τη χρονική στιγμή *t*. Έτσι, οι δυναμικές του περιβάλλοντος καθορίζονται απλά με την εξίσωση:

$$\mathbf{Pr} = \{s_{t+1} = s', r_{t+1} = r | s_t, a_t\} \quad (6)$$

για κάθε s_0, r και για όλες τις πιθανές τιμές των παρελθόντων γεγονότων: s_0, r, s_t και a_t.

Επομένως, ένα σήμα κατάστασης θεωρείται ότι πληροί την ιδιότητα Markov εάν, και μόνον εάν, οι παραπάνω δύο εξισώσεις (5 και 6) ισχύουν για κάθε δυνατή ακολουθία παρελθόντων συμβάντων s_0, r και οποιοδήποτε ιστορικό s_t, a_t, r_t, ..., r_1, s_0, a_0. Σε αυτήν την περίπτωση, τόσο το περιβάλλον όσο και η διαδικασία μάθησης θεωρούνται ότι είναι Markov. Για ένα περιβάλλον Markov, δεδομένης της τρέχουσας κατάστασης και ενέργειας του πράκτορα, με βάση τις δυναμικές ενός βήματος (one-step dynamics), μπορεί να προβλεφθεί η επόμενη κατάσταση και η επόμενη αναμενόμενη ανταμοιβή. Χρησιμοποιώντας επανελλημένα την εξίσωση (6), είναι δυνατόν να προβλεφθούν όλες οι μελλοντικές καταστάσεις και αναμενόμενες ανταμοιβές, σαν να είχε γνωστό όλο το ιστορικό μέχρι εκείνη τη στιγμή. Επομένως, οι καταστάσεις Markov παρέχουν την καλύτερη δυνατή βάση για τη λήψη αποφάσεων. Επιπλέον, μια βέλτιστη πολιτική βασισμένη σε καταστάσεις Markov είναι εξίσου καλή με μια βέλτιστη πολιτική βασισμένη σε πλήρες ιστορικό αλληλεπιδράσεων. Ακόμη και αν το σήμα κατάστασης σε ένα πρόβλημα Ενισχυτικής Μάθησης δεν είναι αυστηρά Markov, συχνά θεωρείται ότι πλησιάζει την ιδιότητα Markov. Σε τέτοιες περιπτώσεις, όσο πιο κοντά πλησιάζει το σήμα στις ιδιότητες ενός Markov σήματος, τόσο μεγαλύτερη είναι η πιθανότητα επιτυχούς εφαρμογής των μεθόδων Ενισχυτικής Μάθησης σε αυτό το πεδίο (Wachi & Sui, 2020).

Μια διαδικασία ΕΜ που πληροί την ιδιότητα Markov ονομάζεται Μαρκοβιανή Διαδικασία Απόφασης (ΜΔΑ) ή Markov Decision Process (MDP). Στην περίπτωση που οι χώροι κατάστασης και ενέργειας είναι πεπερασμένοι, μιλάμε για περατή ΜΔΑ. Η συγκεκριμένη μορφή μιας ΜΔΑ ορίζεται από τα εξής στοιχεία:

- Ένα σύνολο καταστάσεων S
- Ένα σύνολο ενεργειών A
- Μια συνάρτηση ανταμοιβής $R : S \times A \rightarrow R$
- Μια συνάρτηση μετάβασης καταστάσεων $P : S \times A \rightarrow \pi(S)$, όπου κάθε μέλος του $\pi(S)$ είναι μια πιθανοτική κατανομή πάνω στο σύνολο S που αντιστοιχίζει πιθανότητες σε καταστάσεις. Με $P^a_{ss'}$ υποδηλώνεται η πιθανότητα μετάβασης από την κατάσταση s στην κατάσταση s' επιλέγοντας την ενέργεια a.

Η συνάρτηση μετάβασης καταστάσεων καθορίζει με στοχαστικό τρόπο την επόμενη κατάσταση του περιβάλλοντος, δεδομένης της τρέχουσας κατάστασης και ενέργειας του πράκτορα. Αντίστοιχα, η

συνάρτηση ανταμοιβής καθορίζει την αναμενόμενη άμεση ανταμοιβή βάσει της τρέχουσας κατάστασης και ενέργειας (Wachi & Sui, 2020).

Δοθέντων μιας κατάστασης *s* και μιας ενέργειας *a*, η πιθανότητα για κάθε διάδοχη κατάσταση δίνεται από τη σχέση:

$$\mathcal{P}^a_{ss'} = \mathbf{Pr}\{s_{t+1} = s' | s_t = s, a_t = a\} \quad (7)$$

Αυτά τα μεγέθη αναφέρονται ως πιθανότητες μετάβασης. Αντίστοιχα, δεδομένων της τρέχουσας κατάστασης *s* και της τρέχουσας ενέργειας *a*, καθώς και της διάδοχης κατάστασης *s'*, η αναμενόμενη τιμή για την επόμενη ανταμοιβή δίνεται από:

$$\mathcal{R}^a_{ss'} = \mathbf{E}\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \quad (8)$$

Αυτά τα μεγέθη $P^a_{ss'}$ και $R^a_{ss'}$ καθορίζουν τα πιο σημαντικά θέματα που αφορούν τις δυναμικές μιας περατής ΜΔΑ (Wachi & Sui, 2020).

2.2.6 Συναρτήσεις Αξίας

Οι περισσότεροι αλγόριθμοι Ενισχυόμενης Μάθησης βασίζονται στην εκτίμηση των συναρτήσεων αξίας. Αυτές οι συναρτήσεις χρησιμοποιούνται για να αξιολογήσουν πόσο επωφελές είναι να βρίσκεται ο πράκτορας σε μια συγκεκριμένη κατάσταση ή πόσο επωφελές είναι να επιλέξει μια συγκεκριμένη ενέργεια σε αυτήν την κατάσταση. Το "πόσο επωφελές" ορίζεται με βάση τις μελλοντικές ανταμοιβές που μπορούν να αναμένονται, ή αλλιώς, την αναμενόμενη απόδοση. Επειδή οι ανταμοιβές που περιμένει ο πράκτορας εξαρτώνται από τις ενέργειες που θα επιλέξει, οι συναρτήσεις αξίας καθορίζονται βάσει συγκεκριμένων πολιτικών. Όπως αναφέρθηκε προηγουμένως, μια πολιτική *π* ορίζεται ως μια αντιστοίχιση από κάθε κατάσταση *s, a ∈ A(s),* προς την πιθανότητα *π(s, a)* επιλογής της ενέργειας *a* όταν ο πράκτορας βρίσκεται στην κατάσταση *s*. Η αξία μιας κατάστασης *s* συμβολίζεται με $V^{\pi}(s)$ και μπορεί να θεωρηθεί ως η αναμενόμενη επιστροφή όταν ξεκινάμε από την *s* και ακολουθούμε την πολιτική *π* στο εξής (Kaiser et al., 2019).

Η αξία μιας κατάστασης για Μαρκοβιανή Διαδικασία Απόφασης (ΜΔΑ) συμβολίζεται με $V^{\pi}(s)$ και ορίζεται τυπικά ως:

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\} \quad (9)$$

όπου η $E_{\pi}\{\ \}$ αναφέρεται στην αναμενόμενη τιμή, λαμβάνοντας υπόψη την πολιτική π, και t αναπαριστά οποιαδήποτε χρονική στιγμή. Η αξία της τερματικής κατάστασης πάντα ορίζεται ως μηδέν, με βάση τον τρόπο που έχει οριστεί ανεπίσημα η συνάρτηση αξίας. Η V^{π} ονομάζεται συνάρτηση αξίας κατάστασης για την πολιτική π (state-value function).

Με αντίστοιχο τρόπο ορίζεται επίσης η αξία της επιλογής μιας ενέργειας a σε μια κατάσταση s, που συμβολίζεται με $Q^{\pi}(s, a)$, ως η αναμενόμενη απόδοση που προκύπτει από την εκκίνηση στην κατάσταση s, την επιλογή της ενέργειας a, και ακολουθώντας την πολιτική π στη συνέχεια:

$$\mathcal{Q}^{\pi}(s, a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\} \quad (10)$$

Επίσης, η Q^{π} αναφέρεται ως η συνάρτηση αξίας της ενέργειας για την πολιτική π (action-value function).

Μια βασική ιδιότητα των συναρτήσεων αξίας, η οποία εκμεταλλεύονται ευρέως οι τεχνικές Ενισχυτικής Μάθησης και του δυναμικού προγραμματισμού, είναι η ικανότητά τους να ικανοποιούν συγκεκριμένες αναδρομικές σχέσεις. Για κάθε πολιτική π και κάθε κατάσταση s, υπάρχει μια συγκεκριμένη σχέση μεταξύ της αξίας της s και των αξιών των πιθανών διαδοχικών καταστάσεών της:

$$V^{\pi}(s) = \sum_{a} \pi(s, a) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}(s')\right] \quad (11)$$

Οι ενέργειες a επιλέγονται από το σύνολο $A(s)$ και οι επόμενες καταστάσεις s' από το S ή το S^{+} στην περίπτωση των επεισοδιακών προβλημάτων. Η εξίσωση (11) ονομάζεται εξίσωση Bellman και αντιπροσωπεύει τη σχέση μεταξύ μιας κατάστασης και των καταστάσεων που την ακολουθούν. Η εξίσωση Bellman λαμβάνει υπόψιν της όλες τις πιθανές εκβάσεις για την επόμενη κατάσταση, αναθέτοντας σε κάθε μια βαρύτητα αντίστοιχη με την πιθανότητα που αυτή μπορεί να προκύψει. Επιπλέον, δηλώνει ότι η αξία της αρχικής κατάστασης πρέπει να είναι ίση με την εκπτωθείσα αξία της αναμενόμενης επόμενης κατάστασης, προσθέτοντας σε αυτήν την αναμενόμενη στην πορεία ανταμοιβή. Η συνάρτηση αξίας V^{π} είναι η μοναδική λύση στην έκφρασή της ως εξίσωση Bellman. Η εξίσωση

Bellman αποτελεί τη βάση για μια πληθώρα μεθόδων υπολογισμού, προσέγγισης και μάθησης της V^{π} (Kaiser et al., 2019).

Η επίλυση ενός προβλήματος ΜΔΑ συνίσταται στην εύρεση μιας πολιτικής που παρέχει μεγάλη ανταμοιβή σε μακροπρόθεσμο χρονικό ορίζοντα. Για περατές ΜΔΑ, η έννοια της βέλτιστης πολιτικής μπορεί να οριστεί με ακρίβεια μέσω των συναρτήσεων αξίας. Μια πολιτική *π* θεωρείται καλύτερη ή τουλάχιστον ίση με μια πολιτική *π'* εάν η αναμενόμενη επιστροφή της είναι μεγαλύτερη ή τουλάχιστον ίση από αυτήν της *π'* για όλες τις καταστάσεις. Σε άλλα λόγια, ισχύει $\pi \geq \pi'$ αν και μόνον αν $V^{\pi}(s) \geq V^{\pi'}(s)$ για κάθε $s \in S$. Υπάρχει πάντα μια βέλτιστη πολιτική. Ακόμα κι αν υπάρχουν περισσότερες από μία βέλτιστες πολιτικές, αυτές όλες συμβολίζονται με π* και έχουν κοινή συνάρτηση αξίας-κατάστασης, η οποία ονομάζεται βέλτιστη συνάρτηση αξίας-κατάστασης και συμβολίζεται με *V*(s)* , ορίζοντας την ως:

$$V^*(s) = \max_{\pi} V^{\pi}(s) \qquad \forall s \in \mathcal{S} \tag{12}$$

Επίσης, οι βέλτιστες πολιτικές έχουν επίσης μια κοινή βέλτιστη συνάρτηση αξίας-ενέργειας (optimal action-value function), η οποία συμβολίζεται με *Q** και ορίζεται με παρόμοιο τρόπο:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \qquad \forall s \in \mathcal{S}, \quad \forall a \in \mathcal{A}(\mathcal{S}) \tag{13}$$

Για το ζεύγος κατάστασης-ενέργειας *(s, a)*, η βέλτιστη συνάρτηση αξίας-ενέργειας υπολογίζει την αναμενόμενη ανταμοιβή σε περίπτωση που επιλεγεί η ενέργεια *a* στην κατάσταση *s* και ακολουθηθεί η βέλτιστη πολιτική *π* στη συνέχεια. Επομένως, η βέλτιστη συνάρτηση αξίας-ενέργειας *Q** μπορεί να εκφραστεί ως προς τη βέλτιστη συνάρτηση αξίας-κατάστασης *V** ως εξής:

$$Q^*(s,a) = E\{r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\} \tag{14}$$

Επιπλέον, η βέλτιστη συνάρτηση αξίας *V** μπορεί να εκφραστεί βάσει της εξίσωσης Bellman (εξίσωση 11) με μια ειδική μορφή που δεν αναφέρεται σε κάποια συγκεκριμένη πολιτική *π*. Αυτή η μορφή ονομάζεται βέλτιστη συνάρτηση αξίας Bellman (Bellman optimality equation) και αντιπροσωπεύει το γεγονός ότι η αξία μιας κατάστασης υπό τη βέλτιστη πολιτική πρέπει να είναι ίση με την αναμενόμενη ανταμοιβή για την καλύτερη δυνατή ενέργεια στην συγκεκριμένη κατάσταση.

$$V^* = \max_{a \in \mathcal{A}(s)} \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^*(s') \right] \quad (15)$$

Αντίστοιχα η βέλτιστη συνάρτηση αξίας για το *Q** είναι:

$$Q^* = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q^*(s', a') \right] \quad (16)$$

Η εξίσωση βέλτιστης αξίας Bellman *V** (εξίσωση 15) ουσιαστικά αποτελεί ένα σύστημα εξισώσεων με τόσες εξισώσεις όσες είναι και οι καταστάσεις, αλλά διαθέτει μία μοναδική λύση. Έτσι, μπορεί να επιλυθεί χρησιμοποιώντας οποιαδήποτε μέθοδο επίλυσης συστημάτων μη γραμμικών εξισώσεων. Με παρόμοιο τρόπο, η *Q** μπορεί επίσης να εκφραστεί ως ένα σύστημα μη γραμμικών εξισώσεων με παρόμοιες ιδιότητες. Μόλις υπολογιστεί η *V**, υπάρχει η δυνατότητα να καθοριστεί μια βέλτιστη πολιτική. Για κάθε κατάσταση *s*, μπορεί να υπάρχει μία ή περισσότερες ενέργειες που μεγιστοποιούν τη βέλτιστη συνάρτηση Bellman, και συνεπώς υπάρχει η δυνατότητα απλά να οριστεί μια πολιτική που αναθέτει μη μηδενικές πιθανότητες μόνο σε αυτές τις ενέργειες ως βέλτιστη πολιτική (Kaiser et al., 2019).

Στην πράξη, αυτή η πολιτική είναι "άπληστη" ως προς την *V**, και ως εκ τούτου είναι η βέλτιστη πολιτική. Η έννοια του "άπληστου" συνδέεται με τη μυωπική τάση να επιλέγονται οι εναλλακτικές επιλογές που φαίνονται αμέσως βέλτιστες σε ένα πρόβλημα. Ωστόσο, σε αυτήν την περίπτωση, η "άπληστη" πολιτική είναι βέλτιστη ακόμα και μακροπρόθεσμα, καθώς μέσω της *V**, η αναμενόμενη μακροπρόθεσμη επιστροφή είναι διαθέσιμη τοπικά και αμέσως σε κάθε κατάσταση.

Όταν η *Q** είναι διαθέσιμη, τα πράγματα γίνονται ακόμα πιο εύκολα, καθώς δεν απαιτείται η προσεκτική αναζήτηση ενός βήματος μπροστά. Για κάθε κατάσταση *s*, αρκεί να εντοπίσουμε την ενέργεια *a* που μεγιστοποιεί την ποσότητα *Q*(s, a)*. Με το επιπλέον κόστος που απαιτείται για την αναπαράσταση της συνάρτησης αξίας ενέργειας, αποκτάται η δυνατότητα για τη βέλτιστη επιλογή ενεργειών, χωρίς να χρειαστεί να υπάρχει γνώση για τις πιθανές διαδοχικές καταστάσεις και τις αξίες τους, ουσιαστικά δηλαδή τις δυναμικές του περιβάλλοντος (Kaiser et al., 2019).

2.3 Μέθοδοι Ενισχυτικής Μάθησης

Σε αυτήν την ενότητα παρουσιάζονται οι βασικές κατηγορίες μεθόδων για την επίλυση του προβλήματος της ενισχυτικής μάθησης. Τέλος, παρέχεται μια συνοπτική ανασκόπηση τους, με σκοπό να διευκρινιστούν οι διαφορές και οι ομοιότητες μεταξύ τους.

2.3.1 Δυναμικός Προγραμματισμός

Ο όρος "δυναμικός προγραμματισμός" αναφέρεται σε αλγορίθμους που χρησιμοποιούνται για τον υπολογισμό των βέλτιστων πολιτικών σε Μοντέλα Πολλαπλών Διαστάσεων (ΜΔΑ), βασιζόμενοι σε ένα πλήρες μοντέλο του περιβάλλοντος. Παρότι οι κλασικοί αλγόριθμοι δυναμικού προγραμματισμού έχουν περιορισμένη πρακτική χρησιμότητα λόγω της υψηλής υπολογιστικής τους πολυπλοκότητας και της ανάγκης για ένα πλήρες μοντέλο του περιβάλλοντος, η θεωρητική τους σημασία είναι σημαντική. Αποτελούν τη βάση για την κατανόηση πολλών πρακτικών μεθόδων Ενισχυτικής Μάθησης (ΕΜ), οι οποίες προσπαθούν να επιτύχουν τον ίδιο στόχο με μικρότερη υπολογιστική πολυπλοκότητα και χωρίς την ανάγκη για πλήρες μοντέλο του περιβάλλοντος (Busoniu et al., 2017).

Οι κλασικοί αλγόριθμοι δυναμικού προγραμματισμού λειτουργούν με τον τρόπο ότι εξερευνούν εξαντλητικά το χώρο των καταστάσεων και εφαρμόζουν πλήρεις διαδικασίες ανανέωσης πίσω-προς-εμπρός για κάθε κατάσταση. Κάθε ανανέωση πίσω-προς-εμπρός ενημερώνει την αξία μιας κατάστασης, λαμβάνοντας υπόψη τις αξίες όλων των πιθανών επόμενων καταστάσεων και τις πιθανότητες εμφάνισής τους. Μέσω αυτής της διαδικασίας, επιτυγχάνεται σύγκλιση σε τιμές που ορίζονται από τις εξισώσεις Bellman.

Βασικοί όροι στον δυναμικό προγραμματισμό είναι η αξιολόγηση πολιτικής και η βελτίωση πολιτικής. Η αξιολόγηση πολιτικής αναφέρεται στον επαναληπτικό υπολογισμό των συναρτήσεων αξίας για μια συγκεκριμένη πολιτική και μπορεί να αποκαλείται και "πρόβλημα πρόβλεψης". Χρησιμοποιώντας την εξίσωση Bellman, είναι δυνατόν να υπολογιστεί η συνάρτηση αξίας κατάστασης υπό μια δεδομένη πολιτική (Busoniu et al., 2017).

Η βελτίωση πολιτικής αναφέρεται στον υπολογισμό μιας βελτιωμένης πολιτικής, δεδομένης της συνάρτησης αξίας για την τρέχουσα πολιτική. Για μια κατάσταση, είναι αρκετό να εξεταστεί εάν η αξία επιλογής μιας ενέργειας είναι μεγαλύτερη από την αξία που προκύπτει από την τρέχουσα πολιτική. Αν αυτό ισχύει, τότε μια νέα πολιτική που επιλέγει αυτήν την ενέργεια θα είναι καλύτερη. Αυτό αντιστοιχεί στο θεώρημα βελτίωσης πολιτικής.

2.3.2 Monte Carlo

Οι μέθοδοι Monte Carlo ανήκουν σε μια κατηγορία αλγορίθμων Ενισχυτικής Μάθησης που στοχεύουν στην εκμάθηση της αξίας συναρτήσεων και των βέλτιστων πολιτικών, χρησιμοποιώντας εμπειρία υπό μορφή δειγμάτων επεισοδίων. Η απλότητα και η συνάφεια με άλλες μεθόδους Ενισχυτικής Μάθησης είναι κεντρικά στοιχεία των μεθόδων Monte Carlo. Σε αντίθεση με τον Δυναμικό Προγραμματισμό, αυτές οι μέθοδοι δεν απαιτούν πλήρη γνώση του περιβάλλοντος. Αν και η ύπαρξη ενός περιβάλλοντος είναι αναγκαία, οι μέθοδοι Monte Carlo απαιτούν μόνο τη δυνατότητα παραγωγής δειγμάτων μεταβάσεων, ενώ απαιτούν δείγματα ακολουθιών καταστάσεων, ενεργειών και ανταμοιβών (Vodopivec, Samothrakis & Ster, 2017).

Οι μέθοδοι Monte Carlo είναι σημαντικές γιατί επιτρέπουν την εκμάθηση βέλτιστης συμπεριφοράς χωρίς προϋποθέσεις γνώσης για το περιβάλλον. Μπορούν να εκτελούνται είτε μέσω απευθείας αλληλεπίδρασης με το περιβάλλον είτε μέσω εξομοιωμένης αλληλεπίδρασης. Μια από τις σημαντικές προκλήσεις για αυτές τις μεθόδους είναι η διατήρηση επαρκούς επιπέδου εξερεύνησης για να μπορέσουν να εξερευνήσουν όλες τις πιθανές ενέργειες και να αποκτήσουν ολοκληρωμένη εικόνα του περιβάλλοντος. Παρόλες τις διαφορές τους με τον Δυναμικό Προγραμματισμό, οι μέθοδοι Monte Carlo μοιράζονται τη βασική ιδέα της εκτίμησης και βελτίωσης της αξίας, με στόχο την επίτευξη βέλτιστου αποτελέσματος (Vodopivec, Samothrakis & Ster, 2017).

2.3.3 Μάθηση Χρονικών Διαφορών

Η μάθηση με χρονικές διαφορές αποτελεί μια καινοτόμος προσέγγιση στον τομέα της Ενισχυτικής Μάθησης, προσδίδοντας στην έρευνα έναν νέο πυρήνα που αναδεικνύει τη συνδυασμένη χρήση των μεθόδων δυναμικού προγραμματισμού και Monte Carlo. Οι μέθοδοι μάθησης με χρονικές διαφορές (TD) συνδυάζουν την ενημέρωση των εκτιμήσεων αξίας με την ανταμοιβή που λαμβάνει ο πράκτορας σε κάθε χρονική στιγμή, επιτρέποντας την εκτίμηση της βέλτιστης πολιτικής.

Οι μέθοδοι TD επιτρέπουν την ανανέωση των εκτιμήσεων αξίας σε κάθε χρονικό βήμα μέσω του σφάλματος χρονικής διαφοράς (TD error), το οποίο υπολογίζεται με βάση τη διαφορά μεταξύ της τρέχουσας εκτίμησης και της εκτίμησης που προκύπτει μετά από την εφαρμογή της πολιτικής. Αυτή η διαδικασία επιτρέπει τη συνεχή αναβάθμιση των εκτιμήσεων, ανεξαρτήτως του εάν έχει επιτευχθεί οριστικό αποτέλεσμα (Petter, Gershman & Meck, 2018).

Οι μέθοδοι TD είναι εύκολες στην υλοποίηση και επιδεικνύουν αντοχή σε περιβάλλοντα με μακρο-

χρόνιες διαδικασίες, καθιστώντας τις πιο ελκυστικές σε σχέση με τις μεθόδους Monte Carlo. Οι αλγόριθμοι TD συχνά περιγράφονται απλά από μια εξίσωση, όπως ο TD(0), και η εύκολη υλοποίησή τους σε λογισμικό τους καθιστά πρακτικά εφαρμόσιμους σε πραγματικά προβλήματα.

2.3.4 Ίχνη Επιλεξιμότητας

Η χρήση ιχνών επιλεξιμότητας σε συνδυασμό με τα σφάλματα χρονικών διαφορών παρέχει έναν αποδοτικό και επαυξητικό τρόπο προσαρμογής των χαρακτηριστικών των μεθόδων Ενισχυτικής Μάθησης (ΕΜ). Με αυτόν τον τρόπο, οι μέθοδοι ΕΜ μπορούν να κλιμακωθούν και να καλύψουν το εύρος της φιλοσοφίας για τη διαδικασία ενημέρωσης των εκτιμήσεων των συναρτήσεων αξίας, από την ενημέρωση επεισοδίου προς επεισόδιο, των μεθόδων Monte Carlo, έως την ενημέρωση βήμα προς βήμα των μεθόδων Μάθησης Χρονικών Διαφορών (ΜΧΔ) (van Hasselt et al., 2021).

Ενσωματώνοντας ίχνη επιλεξιμότητας στις μεθόδους ΜΧΔ, αυτές αποκτούν χαρακτηριστικά των μεθόδων Monte Carlo, διατηρώντας παράλληλα τα πλεονεκτήματά τους. Η επαυξημένη ευελιξία τους επιτρέπει να αντιμετωπίζουν περιπτώσεις όπου η διαδικασία προς μάθηση δεν είναι πλήρως Markov ή χαρακτηρίζεται από μακροπρόθεσμες ανταμοιβές που καθυστερούν να εμφανιστούν.

Για την εφαρμογή των ιχνών επιλεξιμότητας, κάθε κατάσταση ή ζεύγος κατάστασης-ενέργειας συσχετίζεται με ένα ίχνος επιλεξιμότητας. Σε κάθε χρονικό βήμα, όλα τα ίχνη επιλεξιμότητας φθίνουν κατά παράγοντα, εκτός από το ίχνος της πρόσφατα επισκεφθείσας κατάστασης που αυξάνεται κατά 1. Αυτό επιτρέπει στα ίχνη επιλεξιμότητας να καταγράφουν ποιες καταστάσεις έχει επισκεφθεί πρόσφατα ο πράκτορας.

Στις μεθόδους ΜΧΔ, τα ίχνη επιλογής χρησιμοποιούνται κατά την ανανέωση των αξιών των καταστάσεων. Το σφάλμα χρονικών διαφορών υπολογίζεται ανάλογα με την προηγούμενη κατάσταση, με βάση το ίχνος επιλεξιμότητας. Οι ενημερώσεις αυτές μπορούν να γίνονται σε κάθε βήμα για τους online αλγορίθμους ή στο τέλος του επεισοδίου για off-line αλγορίθμους (van Hasselt et al., 2021).

2.3.5 Μέθοδοι βασισμένοι σε μοντέλο

Κάποιος μπορεί να κατανοήσει υπό μια ενοποιημένη οπτική γωνία τις μεθόδους που χρειάζονται μοντέλα του περιβάλλοντος, όπως ο δυναμικός προγραμματισμός και η ευριστική αναζήτηση, και εκείνες που δεν το απαιτούν, όπως ο Monte Carlo και η μάθηση χρονικών διαφορών. Οι πρώτες

μπορούν να θεωρηθούν ως μέθοδοι σχεδιασμού ενεργειών, ενώ οι δεύτερες ως μέθοδοι μάθησης. Άλλες προσεγγίσεις ονομάζουν τις πρώτες "έμμεση ΕΜ" και τις δεύτερες "άμεση ΕΜ". Η ομοιότητα μεταξύ του σχεδιασμού ενεργειών και της μάθησης για την επίτευξη βέλτιστων συμπεριφορών είναι μεγάλη. Και στις δύο περιπτώσεις, εκτιμούνται οι ίδιες συναρτήσεις αξίας, με επαυξητική ενημέρωση κατά τη διάρκεια μιας ακολουθίας λειτουργιών οπίσθιας ενημέρωσης μικρής κλίμακας. Όλες οι μέθοδοι βασίζονται στην πρόβλεψη των μελλοντικών γεγονότων, τον υπολογισμό αξιών μέσω οπίσθιων ενημερώσεων και τη χρήση τους για την ενημέρωση της εκτιμώμενης συνάρτησης αξίας. Επομένως, οι διαδικασίες μάθησης και σχεδιασμού ενεργειών μπορούν να συνδυαστούν, επιτρέποντας και στις δύο να ενημερώνουν την ίδια εκτιμώμενη συνάρτηση αξίας (Kaiser et al., 2019).

Η διαδικασία αυτή εκμεταλλεύεται την εμπειρία του περιβάλλοντος με δύο διαφορετικούς τρόπους: Μέσω της έμμεσης Ενισχυτικής Μάθησης, μαθαίνουμε και βελτιώνουμε το μοντέλο του περιβάλλοντος, ενώ μέσω της άμεσης Ενισχυτικής Μάθησης, βελτιώνουμε απευθείας τις συναρτήσεις αξίας και την πολιτική. Επιπλέον, μια μέθοδος μάθησης μπορεί να μετατραπεί σε μέθοδο σχεδιασμού ενεργειών, εάν εφαρμοστεί σε εξομοιωμένη εμπειρία αντί της πραγματικής, κλείνοντας ακόμη περισσότερο το χάσμα μεταξύ τους. Αυτό μπορεί να οδηγήσει σε καλύτερα αποτελέσματα με λιγότερες αλληλεπιδράσεις, αν και η ποιότητα του μοντέλου παραμένει κρίσιμη.

Οι διεργασίες μάθησης και σχεδιασμού ενεργειών μπορούν να συνδυαστούν φυσικά, επιτρέποντας στον αλγόριθμο να ενημερώνει την ίδια εκτιμώμενη συνάρτηση αξίας. Αυτή η ομοιότητα επιτρέπει στις δύο διαδικασίες να επωφεληθούν από την αλληλεπίδραση με το περιβάλλον με διαφορετικούς τρόπους:

- Μέσω της έμμεσης Ενισχυτικής Μάθησης, πραγματοποιείται η εκμάθηση και η βελτίωση του μοντέλου του περιβάλλοντος.
- Μέσω της άμεσης Ενισχυτικής Μάθησης, επιτελείται άμεση βελτίωση των συναρτήσεων αξίας και της πολιτικής.

Μια επιπλέον παρατήρηση είναι ότι οποιαδήποτε μέθοδος μάθησης μπορεί να μετατραπεί σε μέθοδο σχεδιασμού ενεργειών αν εφαρμοστεί σε εξομοιωμένη εμπειρία, πράγματι από ένα μοντέλο. Με αυτόν τον τρόπο, το χάσμα μεταξύ των δύο προσεγγίσεων μικραίνει ακόμη περισσότερο, με τη μοναδική διαφορά να είναι η πηγή της εμπειρίας (Kaiser et al., 2019).

Συνολικά, οι διαδικασίες αυτές επιτρέπουν στο σύστημα να μάθει και να βελτιώσει τις συναρτήσεις αξίας και την πολιτική του, ενώ ταυτόχρονα εκτελεί την ενημέρωσή του με διαφορετικούς τρόπους και από διαφορετικές πηγές εμπειρίας.

2.3.6 Ενοποιημένη Άποψη των μεθόδων

Οι προηγούμενοι αλγόριθμοι που αναφέρθηκαν μοιράζονται τρεις βασικές κοινές ιδέες (Lanctot et al., 2017):

1. Στόχος Εκτίμησης Συναρτήσεων Αξίας: Ο κύριος στόχος τους είναι να εκτιμήσουν τις συναρτήσεις αξίας των καταστάσεων ή των καταστάσεων-ενεργειών.

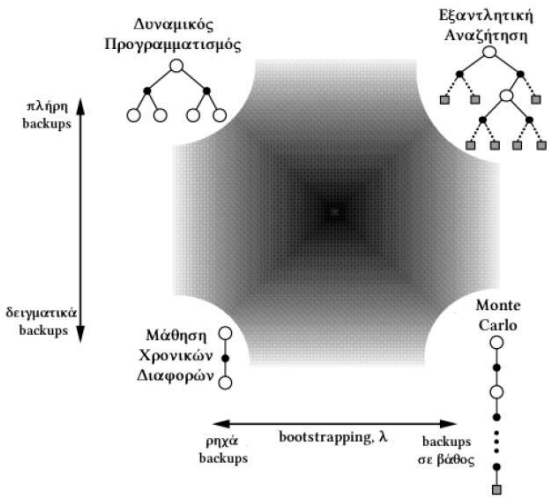
2. Ενημέρωση μέσω Αλληλουχιών Καταστάσεων-Ενεργειών: Όλοι λειτουργούν ενημερώνοντας τιμές βάσει πιθανών ή πραγματικών αλληλουχιών καταστάσεων-ενεργειών.

3. Γενικευμένη Επανάληψη Πολιτικής: Ακολουθούν τη γενικευμένη επανάληψη ως προς την πολιτική, σημαίνοντας ότι διατηρούν μια συνάρτηση αξίας και μια πολιτική κατά προσέγγιση, τις οποίες συνεχώς προσπαθούν να βελτιώσουν μεταξύ τους.

Ωστόσο, υπάρχουν και διαφορές μεταξύ των αλγορίθμων, κυρίως σχετικά με τον τρόπο με τον οποίο ενημερώνονται οι εκτιμήσεις των αξιών. Στην Εικόνα 1, παρουσιάζεται μια αναπαράσταση αυτών των διαφορών:

- Στον κατακόρυφο άξονα, απεικονίζεται η κλίμακα του τρόπου με τον οποίο γίνονται οι ενημερώσεις, από τις δειγματοληπτικές ενημερώσεις έως τις πλήρεις ενημερώσεις του δυναμικού προγραμματισμού.

- Στον οριζόντιο άξονα αντιστοιχίζεται το βάθος στο οποίο φτάνουν οι ενημερώσεις, δηλαδή η βαθμολογία του bootstrapping.



Εικόνα 1: Ενοποιημένη επισκόπηση των μεθόδων ενισχυτικής μάθησης (Πηγή: Lanctot et al., 2017).

2.4 Γενίκευση-Προσέγγιση Συναρτήσεων

Τα συστήματα Ενισχυτικής Μάθησης, ειδικά σε μεγάλη κλίμακα εφαρμογών τεχνητής νοημοσύνης, είναι χρήσιμο να διαθέτουν ικανότητες γενίκευσης. Αυτό συνήθως επιτυγχάνεται μέσω μεθόδων επιβλεπόμενης μάθησης για την προσέγγιση συναρτήσεων, με κάθε ανάστροφη ενημέρωση να θεωρείται ως παράδειγμα εκπαίδευσης για τη συνάρτηση αξίας. Ειδικότερα, οι μέθοδοι βαθμωτής καθόδου κατά την κλίση επιτρέπουν τη φυσική επέκταση με δυνατότητες προσέγγισης συναρτήσεων, συμπεριλαμβανομένων των τεχνικών που συζητήθηκαν προηγουμένως (Liu, Viano & Cevher, 2022).

Ειδικότερα για τις μεθόδους γραμμικής καθόδου κατά την κλίση, υπάρχει μεγάλο ενδιαφέρον, καθώς αποδίδουν καλά και στην πράξη όταν τροφοδοτούνται με τα κατάλληλα χαρακτηριστικά κατάστασης. Η επιλογή των κατάλληλων χαρακτηριστικών κατάστασης είναι κρίσιμη και αποτελεί σημαντικό τρόπο προσθήκης προηγούμενης γνώσης σε συστήματα ΕΜ. Οι μέθοδοι γραμμικής καθόδου κατά την κλίση περιλαμβάνουν τις συναρτήσεις ακτινικής βάσης, την κωδικοποίηση πλακιδίων και την κωδικοποίηση Kanerva. Επίσης δημοφιλείς είναι και οι μέθοδοι ανάστροφης διάδοσης σφάλματος με χρήση νευρωνικών δικτύων, οι οποίες παρουσιάζουν καλές επιδόσεις σε ορισμένες εφαρμογές (Liu, Viano & Cevher, 2022).

Ωστόσο, οι μέθοδοι αυτές εμφανίζουν προβλήματα, όπως το φαινόμενο "ξεμάθαινε", και οι θεωρητικές εγγυήσεις για σύγκλιση σε βέλτιστες πολιτικές είναι πιο αδύναμες σε σχέση με τις γραμμικές μεθόδους.

2.5 Ιεραρχική Ενισχυτική Μάθηση

Όπως συμβαίνει με πολλές μεθόδους και αλγορίθμους τεχνητής νοημοσύνης (και γενικότερα στον τομέα της επιστήμης υπολογιστών), η Ενισχυτική Μάθηση αντιμετωπίζει το πρόβλημα που αποκαλείται "κατάρα της διαστασιμότητας". Αυτό συμβαίνει επειδή ο αριθμός των παραμέτρων που πρέπει να μάθει ο αλγόριθμος αυξάνεται εκθετικά σε σχέση με τον αριθμό των διαφορετικών καταστάσεων που μπορεί να αντιμετωπίσει. Για να αντιμετωπιστεί αυτή η κατάσταση, οι ερευνητές αναζήτησαν τρόπους να εκμεταλλευτούν την έννοια της αφαιρετικότητας ως προς τον χρόνο (Pateria et al., 2021).

Συγκεκριμένα, αντί να λαμβάνουν αποφάσεις σε κάθε χρονική στιγμή, μπορούν να καλούν εκτεταμένες χρονικά δραστηριότητες που ακολουθούν προκαθορισμένες πολιτικές μέχρι να ολοκληρωθούν. Αυτή η προσέγγιση οδηγεί φυσικά σε ιεραρχικές αρχιτεκτονικές και αλγόριθμους μάθησης.

Η χρήση διαφόρων ειδών αφαιρετικότητας έχει βοηθήσει στη διαχείριση προβλημάτων όπως ο σχεδιασμός ενεργειών και η επίλυση προβλημάτων μεγάλης κλίμακας. Αυτό επιτρέπει στο σύστημα να αγνοήσει λεπτομέρειες που δεν είναι σημαντικές για το πρόβλημα. Μια από τις απλούστερες μορφές αφαιρετικότητας είναι η χρήση μακροενεργειών, γνωστών και ως "macros". Ουσιαστικά, αυτές οι macros είναι ακολουθίες ενεργειών που μπορούν να κληθούν με το όνομά τους, σαν να ήταν πρωταρχικές ενέργειες. Αυτό επιτρέπει τη σύνθεση ιεραρχικών ακολουθιών ενεργειών, δίνοντας τη δυνατότητα ακόμα και για την κλήση άλλων macros μέσα στον ορισμό τους. Παρόμοια, υπάρχει η έννοια της υπορουτίνας, όπου εκτός από τις βασικές εντολές επιτρέπεται και η κλήση άλλων υπορουτινών. Οι περισσότερες εργασίες σχετικά με την ιεραρχική ενισχυτική μάθηση χρησιμοποιούν αυτήν την ίδια σημειολογία για την περιγραφή των ιεραρχιών, είτε ως macros είτε ως υπορουτίνες (Pateria et al., 2021).

Από την άποψη της θεωρίας αυτόματου ελέγχου, ένα macro είναι ουσιαστικά μια πολιτική ελέγχου ανοικτού βρόγχου. Ωστόσο, αυτή η προσέγγιση είναι ακατάλληλη για τον έλεγχο σε ένα στοχαστικό σύστημα. Οι ιεραρχικές προσεγγίσεις στην ενισχυτική μάθηση γενικεύουν την έννοια του macro σε πολιτικές ελέγχου κλειστού βρόγχου, οι οποίες ισχύουν για ένα υποσύνολο των καταστάσεων. Αυτές οι μερικές πολιτικές πρέπει να περιλαμβάνουν συγκεκριμένες συνθήκες τερματισμού και μπορούν να αναφερθούν στη βιβλιογραφία ως χρονικά εκτεταμένες ενέργειες, επιλογές, δεξιότητες, συμπεριφορές, τρόποι ή δραστηριότητες.

2.5.1 Ημι-Μαρκοβιανές διαδικασίες απόφασης

Στις Μάρκοβιανες Διαδικασίες Απόφασης (ΜΔΑ), δεν έχει σημασία η χρονική διάρκεια ανάμεσα σε δύο χρονικά βήματα λήψης απόφασης, αλλά η σειριακή φύση της διαδικασίας λήψης απόφασης είναι το κύριο μέλημα. Μια εκδοχή των Μάρκοβιανων Διαδικασιών Απόφασης (ΜΔΑ) είναι οι ημι-Μάρκοβιανες Διαδικασίες Απόφασης (ΗΜΔΑ - Semi-Markov Decision Processes, SMDPs), όπου το διάστημα χρόνου μεταξύ δύο αποφάσεων μπορεί να οριστεί ως (Swishchuk & Vadori, 2017):

- Τυχαία μεταβλητή
- Ακέραια σταθερά
- Πραγματική σταθερά

Στην περίπτωση που ορίζεται ως πραγματική σταθερά, οι ΗΜΔΑ μοντελοποιούν συστήματα διακριτών γεγονότων συνεχούς χρόνου (continuous-time discrete-event systems). Αν οριστεί ως ακέραια

σταθερά, έχουμε τις λεγόμενες διακριτές ΗΜΔΑ (discrete MDPs), όπου οι αποφάσεις λαμβάνονται μόνο σε θετικά ακέραια πολλαπλάσια ενός ορισμένου βασικού χρονικού βήματος.

Σε και τις δύο περιπτώσεις, συνήθως υποθέτουμε ότι το σύστημα παραμένει σε μια κατάσταση για ένα τυχαίο χρονικό διάστημα αναμονής, και κατά τη λήξη αυτού πραγματοποιείται μια ακαριαία μετάβαση προς την επόμενη κατάσταση. Η τυποποίηση των διακριτών ΗΜΔΑ χρησιμοποιείται ευρέως στην ενισχυτική μάθηση, ωστόσο για την επέκτασή της σε περιβάλλοντα με συνεχή χρόνο δεν υπάρχουν σημαντικά θεωρητικά εμπόδια (Swishchuk & Vadori, 2017):

Η επέκταση των ΜΔΑ σε ΗΜΔΑ προσθέτει στο σύνολο των παραδεκτών ενεργειών σύνολα δραστηριοτήτων, επιτρέποντας τον ιεραρχικό καθορισμό μιας καθολικής πολιτικής. Οι αρχικές ενέργειες ονομάζονται θεμελιώδεις και μπορούν να είναι επιλέξιμες ή όχι. Με αυτές τις επεκτάσεις, η διαδικασία λήψης αποφάσεων μοντελοποιείται ως ΗΜΔΑ, όπου ο χρόνος αναμονής σε μια κατάσταση αντιστοιχεί στη χρονική διάρκεια της επιλεγμένης δραστηριότητας. Έτσι, αν *t* είναι ο χρόνος αναμονής στην κατάσταση *s* κατά την εκτέλεση της δραστηριότητας *a*, τότε η *a* διαρκεί *t* βήματα για να ολοκληρωθεί, όταν ξεκινά να εκτελείται στην *s*. Η κατανομή του *t* εξαρτάται από τις πολιτικές και τις συνθήκες τερματισμού όλων των δραστηριοτήτων κατώτερου επιπέδου, που συνθέτουν την *a*.

2.5.2 Προσεγγίσεις

Επιλογές

Μια προσέγγιση για την ιεραρχική Ενισχυτική Μάθηση αποτελούν οι επιλογές (options). Οι επιλογές ορίζονται ως πολιτικές κλειστού βρόγχου, οι οποίες επιλέγουν ενέργειες εντός ενός συγκεκριμένου χρονικού διαστήματος. Παραδείγματα επιλογών μπορεί να είναι δραστηριότητες όπως το ταξίδι προς κάποια απόμακρη πόλη, το άνοιγμα μιας πόρτας ή το δέσιμο των κορδονιών ενός παπουτσιού, καθώς και θεμελιώδεις ενέργειες όπως η συστολή ενός μυ ή το λύγισμα μιας άρθρωσης. Η χρήση των επιλογών επιτρέπει την ένταξη χρονικά αφαιρεμένης γνώσης και δράσης στο πλαίσιο της ενισχυτικής μάθησης με έναν απλό και γενικό τρόπο, ενώ παράλληλα διατηρεί τις ελάχιστες δυνατές αλλαγές στο κλασικό πλαίσιο της ενισχυτικής μάθησης. Συγκεκριμένα, οι επιλογές μπορούν να χρησιμοποιηθούν εναλλασσόμενα με θεμελιώδεις ενέργειες σε μεθόδους σχεδιασμού ενεργειών, όπως ο δυναμικός προγραμματισμός, και σε μεθόδους μάθησης, όπως ο Q-Learning. Το θεωρητικό υπόβαθρο για την προσέγγιση των επιλογών παρέχεται από τις ΗΜΔΑ, όπως έχει ήδη αναφερθεί. Ωστόσο, ένα κρίσιμο χαρακτηριστικό της προσέγγισης είναι η σχέση αλληλεπίδρασης μεταξύ της υποκείμενης ΜΔΑ και της ΗΜΔΑ αυτής καθ' αυτής, καθώς (Lanctot et al., 2019):

- Τα αποτελέσματα του σχεδιασμού ενεργειών με επιλογές μπορούν να χρησιμοποιηθούν κατά την εκτέλεση για τη διακοπή της εκτέλεσης επιλογών, με αποτέλεσμα να επιτυγχάνονται καλύτερα αποτελέσματα από αυτά που είχαν σχεδιαστεί.

- Έχουν προταθεί ενδοεπιλογικές (intra-option) μέθοδοι που μπορούν να μάθουν σχετικά με μια επιλογή με βάση την κατάτμηση της ακολουθίας εκτέλεσής της.

- Μέσω του ορισμού υποστόχων είναι δυνατή η βελτίωση των επιλογών αυτών καθ' εαυτών.

Τέλος, ένα ακόμη σημαντικό χαρακτηριστικό του πλαισίου των επιλογών είναι ότι όλα τα παραπάνω ισχύουν χωρίς να απαιτείται (ή να απαγορεύεται) η χρήση κάποιας συγκεκριμένης μεθοδολογίας αφαίρεσης των καταστάσεων, ιεραρχίας ή συναρτήσεων προσέγγισης.

Ιεραρχίες αφηρημένων μηχανών

Ιεραρχίες Αφηρημένων Μηχανών (Hierarchies of Abstract Machines - HAM) αντιπροσωπεύουν μια προσέγγιση ιεραρχικής οργάνωσης πολιτικών για Μηχανές Διαδικασίας Απόφασης (ΜΔΑ). Παρόμοια με την προσέγγιση των επιλογών, αυτή η μέθοδος βασίζεται στις ΗΜΔΑ, όμως η εστίαση εδώ είναι στην απλοποίηση πολύπλοκων ΜΔΑ με τον περιορισμό των δυνατών πολιτικών, παρά στην αύξηση των διαθέσιμων εναλλακτικών ενεργειών. Συγκεκριμένα, αυτή η προσέγγιση προτείνει την οργάνωση των πολιτικών ως ιεραρχίες στοχαστικών μηχανών πεπερασμένων καταστάσεων. Η βασική ιδέα είναι ότι οι πολιτικές μιας κεντρικής ΜΔΑ μπορούν να περιγραφούν ως προγράμματα που λειτουργούν με βάση τις δικές τους καταστάσεις, εκτός από τις τρέχουσες καταστάσεις της κεντρικής ΜΔΑ. Αυτή η προσέγγιση επιτρέπει τη χρήση προϋπάρχουσας γνώσης για την αποτελεσματική μείωση του χώρου των καταστάσεων, καθώς και τη δημιουργία ενός πλαισίου γνώσης που μπορεί να μεταφερθεί ανάμεσα σε προβλήματα, όπου οι λύσεις μπορούν να συνδυαστούν για την αντιμετώπιση μεγαλύτερων και πιο πολύπλοκων προβλημάτων. Επιπλέον, έχει προταθεί η προσέγγιση των προγραμματιζόμενων HAM (PHAMs), η οποία επεκτείνει τις δυνατότητες των HAM με σκοπό τη βελτίωση της εκφραστικότητάς τους και, πιθανώς στο μέλλον, θα μπορούσε να οδηγήσει στην ανάπτυξη μεθόδων με θεωρητικό υπόβαθρο τη θεωρία βέλτιστου ελέγχου και τη χρήση εκφραστικών γλωσσών προγραμματισμού, προκειμένου να παρέχεται ένα πλήρες και γνωστικά πλούσιο πλαίσιο για την ιεραρχική ενισχυτική μάθηση. Η ολοκλήρωση χαρακτηριστικών όπως οι διακοπές, οι τοπικές μεταβλητές και η παράμετρος πέρασμα σε υπορουτίνες επιτρέπει την εύκολη ενσωμάτωση προϋπάρχουσας γνώσης με συνοπτικό τρόπο (Lanctot et al., 2019).

Ανάλυση Κατάστασης Μεθόδου MAXQ

Μια άλλη προσέγγιση στην ιεραρχική Ενισχυτική Μάθηση είναι η μέθοδος Ανάλυσης Κατάστασης MAXQ (MAXQ Value Function Decomposition). Όπως και οι προηγούμενες προσεγγίσεις των επιλογών και των Ιεραρχιών Αφηρημένων Μηχανών (HAMs), η μέθοδος MAXQ βασίζεται στη θεωρία των Αφηρημένων Μηχανών Ενισχυτικής Μάθησης (ΗΜΔΑ). Ωστόσο, σε αντίθεση με αυτές, η μέθοδος MAXQ δεν εστιάζει απευθείας στη μείωση και απλοποίηση του προβλήματος σε μία ΑΜΕΝ, αλλά σε μια ιεραρχία από ΑΜΕΝ, οι λύσεις των οποίων μπορούν να μαθαίνονται ταυτόχρονα. Μέσω της ανάλυσης της κεντρικής ΜΔΑ σε μικρότερες ΜΔΑ, αποδομείται επίσης η συνάρτηση αξίας σε έναν συνδυασμό των συναρτήσεων αξίας των μικρότερων ΜΔΑ. Αυτή η ανάλυση έχει και μια διαδικασιακή διάσταση, ως μια ιεραρχία υποεργασιών, και μια δηλωτική διάσταση, ως μια αναπαράσταση της συνάρτησης αξίας κατάστασης μιας ιεραρχικής πολιτικής. Η βασική ιδέα είναι ότι ο προγραμματιστής μπορεί να ορίσει χρήσιμους υποστόχους και να ορίσει υποεργασίες που να τους επιτυγχάνουν, μειώνοντας έτσι τον αριθμό των πολιτικών που πρέπει να ληφθούν υπόψη. Η ανάλυση MAXQ είναι σε θέση να αναπαραστήσει οποιαδήποτε πολιτική έχει οριστεί σύμφωνα με μια δοθείσα ιεραρχία. Επιπλέον, δημιουργεί ευκαιρίες για την εκμετάλλευση αφαιρετικών μοντέλων καταστάσεων, επιτρέποντας στις συγκεκριμένες ΜΔΑ μέσα στην ιεραρχία να αγνοούν μεγάλα τμήματα του συνολικού χώρου καταστάσεων. Η MAXQ ξεκινά με μια ανάλυση της κεντρικής ΜΔΑ σε ένα υποσύνολο υποεργασιών $\{M_0, M_1, \ldots, M_n\}$. Αυτές οι υποεργασίες ορίζουν μια ιεραρχία με την M0 να είναι η υποεργασία-ρίζα, που σημαίνει ότι η επίλυσή της συνεπάγεται επίσης και την επίλυση της Μ. Οι ενέργειες που επιλέγονται για την επίλυση της M0 μπορεί να είναι είτε βασικές ενέργειες, είτε πολιτικές που επιλύουν άλλες υποεργασίες, οι οποίες, από την πλευρά τους, μπορούν να αναφέρονται σε βασικές ενέργειες ή πολιτικές για άλλες υποεργασίες. Τέλος, είναι σημαντικό να σημειωθεί ότι η MAXQ μαθαίνει μια αναπαράσταση της συνάρτησης αξίας, με το σημαντικό πλεονέκτημα ότι είναι πιθανή η υλοποίηση και εκτέλεση μιας μη-ιεραρχικής πολιτικής μέσω μιας διαδικασίας παρόμοιας με το βήμα βελτίωσης πολιτικής της διαδικασίας επανάληψης (Lanctot et al., 2019).

2.6 Σχεσιακή Ενισχυτική Μάθηση

Η μεγαλύτερη πλειονότητα των ερευνών σχετικά με την Ενισχυτική Μάθηση βασίζεται στη χρήση προτασιακών αναπαραστάσεων για να περιγράψει τα στοιχεία και τις οντότητες που σχετίζονται με τη διαδικασία μάθησης. Αυτό καθιστά δύσκολη την εφαρμογή των μεθόδων ενισχυτικής μάθησης σε πολύπλοκα πραγματικά προβλήματα. Για την επιτυχή εφαρμογή τεχνικών ενισχυτικής μάθησης

σε προβλήματα υψηλής πολυπλοκότητας, συνήθως απαιτείται η εμπλοκή ειδικευμένων ανθρώπων για τον σχεδιασμό και τη βελτιστοποίηση των προτασιακών αναπαραστάσεων. Ένα κύριο χαρακτηριστικό των προβλημάτων υψηλής πολυπλοκότητας είναι ότι τα χαρακτηριστικά που αφορούν τις καταστάσεις και τις ενέργειες εκφράζονται σε σχέσιακή μορφή και χαρακτηρίζονται από κάποια δομή. Έτσι, η κύρια πρόκληση για ένα σύστημα Τεχνητής Νοημοσύνης που θέλει να επιλύσει αποτελεσματικά τέτοια προβλήματα είναι η αποτελεσματική χρήση των σχεσιακών αυτών δομών για τη μάθηση και τη γενίκευση. Η επιστήμη της Σχεσιακής Ενισχυτικής Μάθησης έχει ως στόχο την επέκταση των πλαισίων της ενισχυτικής μάθησης με αυτές τις δυνατότητες, με σκοπό την αντιμετώπιση προβλημάτων πραγματικού κόσμου υψηλής πολυπλοκότητας (Martínez, Alenya & Torras, 2017).

2.6.1 Στόχοι

Προβλήματα υψηλής πολυπλοκότητας, όπως η εκτέλεση μιας συνταγής μαγειρικής, παρουσιάζουν σημαντικές προκλήσεις για την Ενισχυτική Μάθηση. Αν και θεωρητικά θα μπορούσαν να περιγραφούν με τις αρχές της ενισχυτικής μάθησης, στην πράξη είναι δύσκολο να επιλυθούν με αυτόν τον τρόπο, καθώς οι πολύπλοκες διαδικασίες αυτές εκφράζονται φυσικότερα μέσω σχέσεων και δομών. Η αποτελεσματική χρήση αυτών των δομών αποτελεί πρόκληση για την Ενισχυτική Μάθση. Στην ενότητα αυτή περιγράφονται εν συντομία οι προκλήσεις που αντιμετωπίζουν οι ερευνητές της Ενισχυτικής Μάθησης, με στόχο την πρόοδο προς αυτήν την κατεύθυνση (Martínez, Alenya & Torras, 2017).

Πρώτον, η ανάπτυξη προσεγγίσεων προσέγγισης συναρτήσεων είναι αναγκαία, καθώς οι υπάρχουσες μέθοδοι δεν είναι ιδανικές για την αναπαράσταση σχέσιακής γνώσης. Δεν παρέχουν επαρκείς δυνατότητες γενίκευσης σε πολύπλοκα προβλήματα, εκτός αν προηγηθεί λεπτομερής μελέτη και σχεδιασμός των χαρακτηριστικών του πεδίου που θα χρησιμοποιηθούν από τη μέθοδο προσέγγισης.

Δεύτερον, είναι επιθυμητό να βελτιωθεί η δυνατότητα μεταφοράς γνώσης μεταξύ παρόμοιων αντικειμένων. Η εντοπισμός παρόμοιων αντικειμένων, για τα οποία υπάρχει πιθανή γενίκευση, αποτελεί δύσκολη διαδικασία.

Τρίτον, η μεταφορά γνώσης μεταξύ διαδικασιών απαιτεί συστηματική προσέγγιση. Ενώ θέλουμε οι πράκτορες να είναι ικανοί για γενικευμένη χρήση σε ένα συγκεκριμένο πεδίο, η πραγματικότητα συχνά αντιμετωπίζει δυσκολίες σε αυτόν τον τομέα.

Τέταρτον, η προσέγγιση στον σχεδιασμό ενεργειών και στην εξαγωγή συμπερασμάτων απαιτεί συνδυασμό συνειδητών αποφάσεων και αντίδρασης. Η προσέγγιση αυτή πρέπει να λαμβάνει υπόψη την προσέγγιση στην αναζήτηση και την ανάγκη για βελτίωση της ακρίβειας στην πρόβλεψη των συναρτήσεων αξίας.

Τέλος, η πρότερη γνώση είναι σημαντική, αλλά συχνά η ενισχυτική μάθηση βασίζεται περισσότερο στη δοκιμή και το σφάλμα για την εκπαίδευση των συστημάτων, κάτι που μπορεί να είναι μη αποδοτικό σε πολύπλοκες διαδικασίες (Martínez, Alenya & Torras, 2017).

2.6.2 Μέθοδοι

Σε αυτήν την ενότητα περιγράφονται επιγραμματικά μερικές ελπιδοφόρες προσεγγίσεις που σχετίζονται με τον τομέα της σχεσιακής ενισχυτικής μάθησης (Martínez, Alenya & Torras, 2017).

Η πρώτη προσέγγιση που αναφέρεται είναι η σχεσιακή παλινδρόμηση σε συνδυασμό με το Q-Learning. Η χρήση σχεσιακής παλινδρόμησης επιτρέπει την εφαρμογή του Q-Learning σε περιβάλλοντα που χαρακτηρίζονται από σχέσεις. Με τη χρήση σχεσιακής αναπαράστασης των καταστάσεων και των ενεργειών, η σχεσιακή παλινδρόμηση επιτρέπει τη γενίκευση των τιμών της συνάρτησης Q, εκμεταλλευόμενη τη δομημένη πληροφορία και την υπάρχουσα εμπειρία σε σχετικά προβλήματα.

Η δεύτερη προσέγγιση είναι η προσεγγιστική επανάληψη πολιτικής, η οποία στηρίζεται στην άμεση αναπαράσταση των πολιτικών σε συνδυασμό με την υποννοούμενη έμμεση αναπαράσταση των συναρτήσεων αξίας. Αυτή η μέθοδος διευκολύνει την εκμάθηση πολιτικών κατάλληλων για δομημένα πεδία, χρησιμοποιώντας γλώσσες γενικού σκοπού που επιτρέπουν τη συμπαγή περιγραφή πολλών χρήσιμων πολιτικών.

Η τρίτη προσέγγιση είναι ο συμβολικός δυναμικός προγραμματισμός, που στοχεύει στην εκμετάλλευση της συμβολικής αναπαράστασης του μοντέλου μετάβασης καταστάσεων. Αυτή η προσέγγιση επιδιώκει τη δημιουργία μιας συμβολικής έκδοσης των συναρτήσεων αξίας με χρήση παλινδρόμησης.

Τέλος, η τέταρτη προσέγγιση είναι η άμεση προσέγγιση των συναρτήσεων αξίας, η οποία στοχεύει στη βελτίωση της αναπαράστασης των συναρτήσεων αξίας ώστε να εκμεταλλεύονται τη σχεσιακή δομή του πεδίου. Αυτή η προσέγγιση χρησιμοποιεί τεχνικές γραμμικού προγραμματισμού για την άμεση προσέγγιση της συνάρτησης αξίας, επιτρέποντας την ανάλυση των τοπικών συναρτήσεων αξίας για κάθε κλάση αντικειμένων (Martínez, Alenya & Torras, 2017).

2.7 Ποιότητα της Διαδικασίας Μάθησης

Εκτός από την αξιολόγηση της πολιτικής που εκμαθαίνει ο πράκτορας, είναι σημαντικό να εξετάζουμε και να εκτιμούμε την ποιότητα της ίδιας της διαδικασίας μάθησης. Για αυτό τον λόγο, χρησιμοποιούνται διάφορες μετρικές, πολλές εκ των οποίων είναι ασύμβατες μεταξύ τους. Αναφέρονται και περιγράφονται σύντομα μερικές από αυτές (Bellemare et al., 2013):

1. **Σύγκλιση στην Οπτική Συμπεριφορά:** Ορισμένοι αλγόριθμοι προσφέρουν θεωρητικές εγγυήσεις για τη σύγκλισή τους στη βέλτιστη συμπεριφορά. Αν και αυτό παρέχει μερικές εγγυήσεις, η πρακτική αξία τους είναι περιορισμένη, καθώς ο πράκτορας που επιτυγχάνει συμπεριφορά που είναι βέλτιστη σε ποσοστό μόνο 99% μπορεί να είναι προτιμότερος από έναν που σιγά-σιγά συγκλίνει στο 100% της βέλτιστης συμπεριφοράς, αλλά με πολύ αργό ρυθμό μάθησης.

2. **Ταχύτητα Σύγκλισης στο Βέλτιστο:** Καθώς η βελτιστότητα συνήθως ορίζεται ασυμπτωτικά, ο ορισμός της ταχύτητας σύγκλισης στο βέλτιστο δεν μπορεί να είναι θεωρητικά ορθός. Ένα πρακτικότερο μέτρο είναι η σύγκλιση κοντά στο βέλτιστο. Ωστόσο, ακόμα και με αυτήν τη μετρική, χρειάζεται να οριστεί πόσο κοντά στη βέλτιστη συμπεριφορά είναι επιθυμητό να φτάσουμε.

3. **Μετάνοια (Regret):** Μια πιο κατάλληλη μετρική για την εκτίμηση της ποιότητας της διαδικασίας μάθησης είναι η μετάνοια. Αυτή ορίζεται ως η αναμενόμενη μείωση της ανταμοιβής που προκύπτει από την εκτέλεση του αλγορίθμου μάθησης, σε σχέση με την εφαρμογή της βέλτιστης πολιτικής από την αρχή. Η αξιολόγηση της μετάνοιας είναι ωστόσο δύσκολη.

3 Μηχανική Μάθηση και Ηλεκτρονικά Παιχνίδια

3.1 Εισαγωγή

Τα παιχνίδια, είτε είναι ηλεκτρονικά είτε όχι, από πάντα αποτελούσαν μια προκλητική δραστηριότητα για τον ανθρώπου, πέραν της απλής διασκέδασης. Αυτό οδήγησε στο να γίνουν ένα από τα πιο δημοφιλή πεδία έρευνας για τους επιστήμονες που ασχολούνται με την τεχνητή νοημοσύνη και τη μηχανική μάθηση. Αρχικά, οι έρευνες επικεντρώθηκαν σε παιχνίδια όπως το τάβλι, το σκάκι, η ντάμα και το πόκερ, όπου επιτεύχθηκαν εξαιρετικά αποτελέσματα. Παραδείγματος χάριν, η εργασία του Arthur Samuel (1959), η οποία ήταν η πρώτη εφαρμογή μηχανικής μάθησης σε αναλυτικά παιχνίδια, επηρέασε σημαντικά την επιστημονική κοινότητα.

Η επιλογή των αναλυτικών παιχνιδιών για έρευνα στη μηχανική μάθηση ήταν λογική, καθώς παρέχουν πλήρη πληροφορία για το περιβάλλον τους και έχουν διαφανείς και περιορισμένους κανόνες. Αντίθετα, τα εμπορικά ηλεκτρονικά παιχνίδια συνήθως εμπλέκουν αλληλεπιδράσεις μεταξύ πολλών αντικειμένων, γεγονός που οδηγεί σε δυσκολίες στην εφαρμογή μηχανικής μάθησης.

Παρά τις αρχικές προκλήσεις, οι ερευνητές σήμερα στρέφουν το ενδιαφέρον τους προς τα ηλεκτρονικά παιχνίδια, καθώς πιστεύουν ότι οι τεχνολογικές εξελίξεις τους επιτρέπουν να αντιμετωπίσουν τις προκλήσεις της τεχνητής νοημοσύνης σε αυτό τον τομέα. Οι εφαρμογές της τεχνητής νοημοσύνης στα ηλεκτρονικά παιχνίδια μπορεί να βελτιώσουν την εμπειρία των παικτών και να παράσχουν ενδιαφέρουσες προκλήσεις για την έρευνα στον τομέα (de Almeida Rocha & Duarte, 2019).

Παρόλα αυτά, μέχρι τώρα η κυρίαρχη τάση στη βιομηχανία ανάπτυξης ηλεκτρονικών παιχνιδιών ήταν η προσπάθεια να προγραμματίσουν όλες τις πιθανές καταστάσεις και αντιδράσεις στο μοντέλο του κόσμου του παιχνιδιού. Αυτό οφείλεται στη σχεδόν απόλυτη προσέγγιση τους στον προγραμματισμό, χωρίς να ενσωματώνουν στοιχεία τεχνητής νοημοσύνης στα παιχνίδια. Εντούτοις, αναμένεται ότι αυτό θα αλλάξει σύντομα, καθώς οι τεχνολογίες γραφικών φθάνουν σε σημείο κορεσμού και παραχωρούν περιθώριο για την ανάπτυξη της τεχνητής νοημοσύνης στα παιχνίδια.

Ταυτόχρονα, οι χρήστες αναζητούν όλο και πιο ρεαλιστικές εμπειρίες παιχνιδιού, πράγμα που απαιτεί εξυπνότερους και πιο ανταγωνιστικούς εικονικούς αντιπάλους. Έτσι, η εφαρμογή προηγμένων τεχνικών τεχνητής νοημοσύνης στα παιχνίδια γίνεται όλο και πιο ελκυστική ως το επόμενο βήμα για τους παραγωγούς και τους παίκτες. Η ένταξη της μηχανικής μάθησης στα υποσυστήματα τεχνητής νοημοσύνης των μοντέρνων παιχνιδιών μπορεί να βελτιώσει το gameplay των παιχνιδιών επόμενης γενιάς. Για παράδειγμα, τα παιχνίδια με δυνατότητες μάθησης μπορούν να βελτιώσουν τη συμπερι-

φορά των εικονικών αντιπάλων, να προσφέρουν εμπειρίες gameplay μέσω "εκπαιδευτικής" διαδικασίας και να προσαρμόζουν τις στρατηγικές τους ανάλογα με τις ενέργειες του παίκτη (de Almeida Rocha & Duarte, 2019).

Στην ενότητα αυτή, εξετάζονται οι προοπτικές που ανοίγονται μέσω της εφαρμογής τεχνικών Μηχανικής Μάθησης σε εμπορικά ηλεκτρονικά παιχνίδια. Αρχικά, αναλύονται ορισμένα θέματα που διέπουν τον σχεδιασμό και την υλοποίηση συστημάτων Τεχνητής Νοημοσύνης για ηλεκτρονικά παιχνίδια, προκειμένου να διευκρινιστούν οι συγκεκριμένες απαιτήσεις και ιδιαιτερότητες του κλάδου. Στη συνέχεια, αναδεικνύονται οι προοπτικές που προκύπτουν από τη χρήση μεθόδων Μηχανικής Μάθησης σε ηλεκτρονικά παιχνίδια. Ακολουθεί μια σύντομη επισκόπηση των διαφόρων τεχνικών Μηχανικής Μάθησης και του πώς αυτές μπορούν να ενσωματωθούν σε συστήματα Τεχνητής Νοημοσύνης ηλεκτρονικών παιχνιδιών. Έπειτα, παρουσιάζονται τρεις σύντομες μελέτες περίπτωσης εμπορικών παιχνιδιών που εφαρμόζουν τεχνικές Μηχανικής Μάθησης. Τέλος, δίνονται κάποια πρώιμα συμπεράσματα και γίνεται προσπάθεια να εκτιμηθούν εν συντομία οι επιπτώσεις μιας πιθανής σύγκλισης μεταξύ της ερευνητικής κοινότητας και της βιομηχανίας ηλεκτρονικών παιχνιδιών, την οποία θεωρούμε ως προϋπόθεση για την επιτάχυνση των εξελίξεων προς την κατεύθυνση που εξετάζεται στην εργασία.

3.2 AI και ηλεκτρονικά παιχνίδια

3.2.1 Η χρησιμότητα της AI

Η Τεχνητή Νοημοσύνη στα ηλεκτρονικά παιχνίδια εκφράζεται κυρίως μέσω των ενεργειών των χαρακτήρων που ελέγχονται από τον υπολογιστή (Non-Playing Characters - NPCs), οι οποίοι μπορούν να θεωρηθούν ως πράκτορες. Στο πεδίο των παιχνιδιών, μπορούμε να αναγνωρίσουμε τρία επίπεδα νοημοσύνης που αφορούν αυτούς τους πράκτορες (από χαμηλότερο σε υψηλότερο επίπεδο) (Millington, 2019):

- **Λειτουργικό (Operational):** Περιλαμβάνει τις απλές κινήσεις και ενέργειες των πρακτόρων σε χαμηλό επίπεδο.
- **Τακτικό (Tactical):** Ορίζει ακολουθίες ενεργειών του πράκτορα για την επίτευξη συγκεκριμένων στόχων στο περιβάλλον δράσης.
- **Στρατηγικό (Strategic):** Αναφέρεται σε μακροπρόθεσμες αποφάσεις σχεδιασμού από τον

πράκτορα.

Για να διακριθεί καλύτερα το τακτικό από το στρατηγικό επίπεδο, είναι σημαντικό να σημειωθεί ότι οι στρατηγικές αποφάσεις καθορίζουν τους στόχους προς τους οποίους στοχεύει ο πράκτορας. Οι ρόλοι των πρακτόρων στα ηλεκτρονικά παιχνίδια μπορεί να είναι οι εξής:

- **Αντίπαλοι:** Δημιουργείται η αίσθηση ότι ο παίκτης αντιμετωπίζει άλλον παίκτη. Απαιτούνται αποφάσεις τόσο τακτικού όσο και στρατηγικού επιπέδου.

- **Σύμμαχοι:** Σκοπός τους είναι να υποστηρίξουν τον παίκτη μέσω συνεργασίας ή παροχής υποδείξεων και συμβουλών.

- **Υποστηρικτικοί Χαρακτήρες:** Έχουν ουδέτερη προδιάθεση έναντι του παίκτη και ο στόχος τους είναι να κάνουν τον εικονικό κόσμο πιο ρεαλιστικό.

Βάσει των προηγούμενων, το πρόβλημα της Τεχνητής Νοημοσύνης στα ηλεκτρονικά παιχνίδια περιγράφεται ως την ανάγκη για ευφυή επιλογή των κατάλληλων αποφάσεων σε κάθε επίπεδο, με στόχο τη δημιουργία της πιο αληθοφανούς συμπεριφοράς από το σύστημα Τεχνητής Νοημοσύνης (που, φυσικά, συνεισφέρει σε βελτιωμένο gameplay). Οι κατασκευαστές παιχνιδιών, σε αντίθεση με τους ακαδημαϊκούς ερευνητές, δεν έχουν ως κύριο στόχο την εύρεση και επιλογή (ακόμα και προσεγγιστικών) βέλτιστων λύσεων ή αποφάσεων από τα συστήματα Τεχνητής Νοημοσύνης που ενσωματώνουν στα παιχνίδια τους. Ο κύριος στόχος τους είναι να συνεισφέρουν στο συνολικό χαρακτήρα του παιχνιδιού, προσφέροντας διασκέδαση. Παρόλα αυτά, ενώ η έρευνα μπορεί να αξιολογηθεί με βάση αξιόπιστες μετρικές, η βιομηχανία εστιάζει σε εμπειρικά κριτήρια για την αξιολόγηση συστημάτων Τεχνητής Νοημοσύνης σε εμπορικά ηλεκτρονικά παιχνίδια. Κατά συνέπεια, η διαδικασία σχεδίασης συστημάτων Τεχνητής Νοημοσύνης για παιχνίδια θα πρέπει να λαμβάνει υπόψη της τα εξής ζητήματα (Millington, 2019):

- Το παιχνίδι ως πρόκληση, καθώς δεν είναι ενδιαφέρον αν είναι πολύ εύκολο ή πολύ δύσκολο.

- Συχνά, οι παίκτες νιώθουν ότι οι ήττες τους είναι άδικες.

- Ένα σύστημα Τεχνητής Νοημοσύνης που μπορεί να διατηρήσει την ψευδαίσθηση της ευφυΐας εκτιμάται θετικά από τον παίκτη.

Για να επιτευχθούν οι προαναφερθέντες στόχοι στο μέγιστο δυνατό βαθμό, είναι απαραίτητο να τηρούνται οι ακόλουθες γενικές αρχές:

- Αποφυγή προφανούς απάτης (cheating), η οποία συνδέεται με το δεύτερο σημείο.
- Μη προβλέψιμη συμπεριφορά, σχετιζόμενη με τα πρώτα δύο σημεία.
- Αποφυγή μη προφανούς ευτελούς συμπεριφοράς, που αφορά τα τρία από τα παραπάνω σημεία.
- Χρήση του περιβάλλοντος.
- Αυτοδιόρθωση, η οποία ανήκει στα πρώτα και τα τρίτα σημεία.
- Δημιουργικότητα, που συνδέεται με το τρίτο σημείο.
- Συμπεριφορά που να μοιάζει με ανθρώπινη, σχετιζόμενη με τα δύο τελευταία σημεία.

Μέχρι στιγμής, φαίνεται ότι οι παραγωγοί παιχνιδιών λαμβάνουν υπόψη τους μόνο τις τρεις πρώτες αρχές. Ωστόσο, οι τεχνικές Μηχανικής Μάθησης θα μπορούσαν να προσφέρουν λύσεις προς αυτή την κατεύθυνση (Millington, 2019).

3.2.2 Κλασσικές Μέθοδοι AI

Η χρήση παραδοσιακών μεθόδων Τεχνητής Νοημοσύνης είναι συνήθης πρακτική για τους κατασκευαστές εμπορικών παιχνιδιών. Αυτές οι μέθοδοι έχουν καθιερωθεί στη βιομηχανία παιχνιδιών ως ένα είδος καθολικού status quo, πράγμα που καθιστά απίθανη την πλήρη αντικατάστασή τους από σύγχρονες τεχνικές Τεχνητής Νοημοσύνης όπως η Μηχανική Μάθηση. Ως εκ τούτου, είναι σημαντικό να αναφερθούν ορισμένες από τις πιο δημοφιλείς κλασικές τεχνικές που χρησιμοποιούνται σήμερα σε εμπορικά ηλεκτρονικά παιχνίδια (Alam, 2022):

- **Απάτη (Cheating):** Σε αυτήν την τεχνική, ο υπολογιστής "απατάει" τον ανθρώπινο αντίπαλο του, χρησιμοποιώντας παράνομα πλεονεκτήματα όπως πλήρης όραση και δυνατότητα διέλευσης μέσα από εμπόδια. Αν και αυτή η τεχνική μπορεί να οδηγήσει σε αποτελέσματα, διακρίνονται τάσεις απομάκρυνσης από αυτήν.
- **Θεωρία Παιγνίων:** Αυτή η τεχνική εφαρμόζεται σε απλά παιχνίδια όπως η τρίλιζα, με εξαιρετικά αποτελέσματα, καθώς έχουν αναπτυχθεί αλγόριθμοι που είναι αδύνατον να νικηθούν.

- **Δένδρα Παιγνίων και Μέθοδοι Αναζήτησης:** Αυτές οι τεχνικές βασίζονται στην υπολογιστική ισχύ για την επίτευξη καλών επιδόσεων σε παιχνίδια όπως το τάβλι, το σκάκι και η ντάμα.

- **Σχεδιασμός Μονοπατιού:** Αυτή η τεχνική, η οποία ήταν δύσκολη στο παρελθόν, τώρα λύνεται κυρίως μέσω του αλγορίθμου A*, εκτός από κάποιες δυσκολίες στις τρισδιάστατες περιπτώσεις.

- **Μέθοδοι βασισμένες σε κανόνες:** Αυτή η μέθοδος είναι ευρέως χρησιμοποιούμενη και σχετίζεται με την εφαρμογή προκαθορισμένων κανόνων.

- **Μηχανές πεπερασμένων καταστάσεων:** Αυτές οι μηχανές χρησιμοποιούνται για τη δομημένη υλοποίηση των συστημάτων Τεχνητής Νοημοσύνης.

- **Flocking:** Αυτή η τεχνική αναφέρεται στην ομαλή και συντονισμένη κίνηση ομάδων οντοτήτων, όπως μονάδες σε παιχνίδια στρατηγικής.

- **Ιεραρχική Τεχνητή Νοημοσύνη:** Αυτή η τεχνική περιλαμβάνει τη δομή του συστήματος Τεχνητής Νοημοσύνης σε δύο επίπεδα και τη χρήση διαφορετικών τεχνικών σε αυτά.

Κοινό στοιχείο όλων αυτών των τεχνικών είναι ότι περιορίζονται στη χρήση προκαθορισμένων κανόνων και αδυνατούν να παράγουν νέες λύσεις για τα προβλήματα που παρουσιάζονται. Αυτό το γεγονός αμφισβητεί την πραγματική ευφυΐα των συστημάτων που βασίζονται σε αυτές τις μεθόδους, καθώς αναμένεται μια ευφυή οντότητα να μπορεί να βελτιώσει τη συμπεριφορά της μαθαίνοντας (Alam, 2022).

3.2.3 Ψευδομάθηση

Πριν εξεταστεί η Μηχανική Μάθηση (ΜΜ), είναι σημαντικό να γίνει αναφορά σε μια τάση που παρατηρείται στη βιομηχανία των ηλεκτρονικών παιχνιδιών: η χρήση διαφόρων τεχνικών για να δημιουργηθεί η ψευδαίσθηση ότι το τμήμα Τεχνητής Νοημοσύνης (ΤΝ) του παιχνιδιού μαθαίνει. Στη βιομηχανία, είναι συνηθισμένο να αναπτύσσονται συστήματα ΤΝ που βασίζονται στην απάτη, καθώς αυτό επιπλέον απλοποιεί την υλοποίησή τους. Αν και αυτή η πρακτική δεν είναι απαραίτητα κακή αν υλοποιηθεί με τρόπο που δεν είναι εύκολα αντιληπτός από τους παίκτες, είναι σημαντικό να είναι υπό διαρκή εξέταση (Frutos-Pascual & Zapirain, 2015).

Μια συνήθης πρακτική για τη δημιουργία της αίσθησης ότι το σύστημα ΤΝ ενός παιχνιδιού μαθαίνει είναι ο προγραμματισμός διαφόρων επιπέδων επίδοσης βάσει ενός προκαθορισμένου συνόλου συμπεριφορών και η δυναμική εναλλαγή των επιπέδων καθώς ο παίκτης προχωρά στο παιχνίδι. Μια πιο δομημένη προσέγγιση είναι η προσθήκη νέων στοιχείων συμπεριφοράς κατά τη διάρκεια του παιχνιδιού, όπως το ξεκλείδωμα νέων καταστάσεων σε μηχανές πεπερασμένων καταστάσεων ή νέων κανόνων σε συστήματα βασισμένα σε κανόνες.

Επιπλέον, η αίσθηση αυτή μπορεί να δημιουργηθεί με την προσαρμογή διαφόρων παραμέτρων κατά τη διάρκεια του παιχνιδιού, όπως ο ρυθμός λαθών, η ακρίβεια, ο χρόνος αντίδρασης και η επιθετικότητα (Frutos-Pascual & Zapirain, 2015).

3.3 Μηχανική Μάθηση και Αναλυτικά Παιχνίδια

Ντάμα

Η εφαρμογή της Μηχανικής Μάθησης και της ενισχυτικής μάθησης σε παιχνίδια έχει ιστορία που ξεκινά από τα τέλη της δεκαετίας του '50, όταν ο Samuel (1959) παρουσίασε ένα πρόγραμμα που είχε τη δυνατότητα να μαθαίνει από μόνο του. Αυτή η εργασία είχε μεγάλη επίδραση στην ερευνητική κοινότητα και θεωρείται από πολλούς ως το πρώτο πρόγραμμα ηλεκτρονικού υπολογιστή που μπορούσε να μαθαίνει αυτόνομα.

Σκάκι

Το σκάκι είναι το πιο δημοφιλές παιχνίδι στην ερευνητική κοινότητα της Τεχνητής Νοημοσύνης. Οι καλύτερες προσεγγίσεις για πράκτορες που παίζουν σκάκι βασίζονται σε δένδρα αναζήτησης παιγνίων, χρήση γνώσης αποθηκευμένης σε βάσεις δεδομένων και υπολογιστική ισχύ. Αν και οι προσεγγίσεις που βασίζονται σε Μηχανική Μάθηση δεν είναι τόσο επιτυχείς όταν παίζουν εναντίον ανθρώπων, παραμένουν ενδιαφέρουσες (Hitar-Garcia et al., 2022).

Go (ιαπωνέζικο επιτραπέζιο)

Το Go είναι γνωστό για την απλότητα των κανόνων του αλλά και για την μεγάλη του πολυπλοκότητα. Οι προσπάθειες στην ερευνητική κοινότητα έχουν εστιαστεί στη μάθηση με βάση συναρτήσεις

αξιολόγησης και στην ανάλυση των χαρακτηριστικών που επηρεάζουν το τελικό αποτέλεσμα.

Τάβλι

Το τάβλι είναι μια επιτυχημένη εφαρμογή της Μηχανικής Μάθησης στα αναλυτικά παιχνίδια. Προγράμματα όπως το NeuroGammon χρησιμοποιούν νευρωνικά δίκτυα για να κερδίσουν σε διαγωνισμούς τάβλι.

Poker

Το poker είναι πρόκληση για την ΤΝ λόγω της ατελούς πληροφορίας. Έχουν προταθεί προσεγγίσεις που βασίζονται στην μοντελοποίηση του αντιπάλου για την ανάλυση των κινήσεων του. Προγράμματα όπως το Loki είναι γνωστά παραδείγματα υλοποιήσεων Μηχανικής Μάθησης στο poker (Hitar-Garcia et al., 2022).

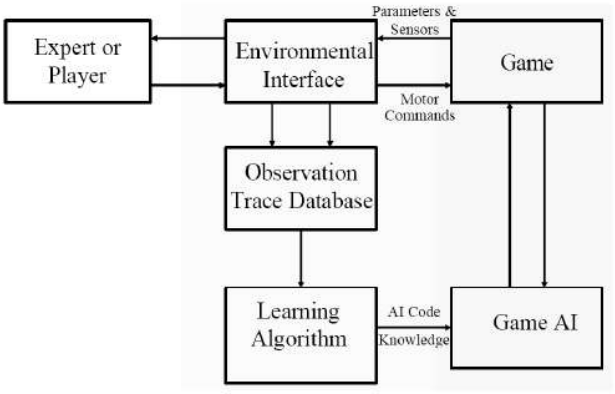
3.4 Ζητήματα εφαρμογής σε εμπορικά παιχνίδια

3.4.1 Πηγές Μάθησης

Σε αυτήν την ενότητα παρουσιάζονται διάφορες προσεγγίσεις για την απόκτηση γνώσης που θα χρησιμοποιηθεί από τις μεθόδους Μηχανικής Μάθησης (Bertolini et al., 2021).

Παρατήρηση Ανθρώπινης Συμπεριφοράς

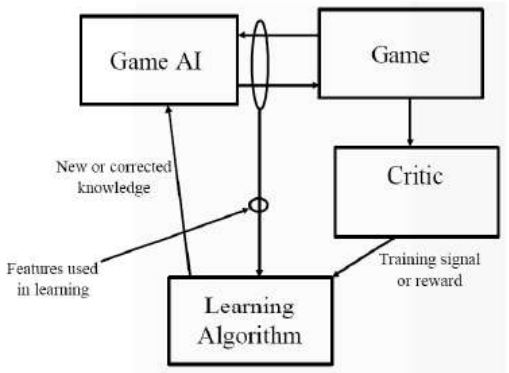
Η μάθηση γίνεται μέσω της παρατήρησης της συμπεριφοράς ανθρώπων και της αναπαραγωγής των συγκεκριμένων συμπεριφορών. Στόχος είναι να αντιγραφούν αυτές οι συμπεριφορές με ακρίβεια, λαμβάνοντας υπόψη τις διαφορές στην προσωπικότητα, τον τρόπο έκφρασης και την κουλτούρα των ανθρώπων που παρατηρούνται.



Εικόνα 2: Μάθηση μέσω παρατήρησης ανθρώπινης συμπεριφοράς (Πηγή: Bertolini et al., 2021).

Καθοδήγηση

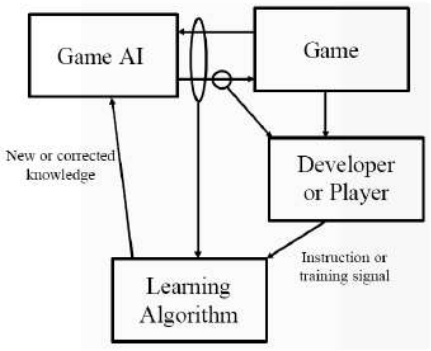
Η εκπαίδευση γίνεται από μη προγραμματιστές, που καθοδηγούν τη μηχανή ΤΝ με οδηγίες προς μια συγκεκριμένη συμπεριφορά-στόχο.



Εικόνα 3: Μάθηση μέσω καθοδήγησης (Πηγή: Bertolini et al., 2021).

Εμπειρία

Η μηχανή ΤΝ αποκτά εμπειρία ανταγωνιζόμενη άλλες μηχανές ΤΝ ή ανθρώπους, κατά τη διάρκεια της ανάπτυξης του παιχνιδιού. Με αυτόν τον τρόπο, βελτιώνεται η συμπεριφορά της μηχανής και ανιχνεύονται κίβδηλες συμπεριφορές. Επιπλέον, η μηχανή μπορεί να εξερευνήσει το περιβάλλον της από μόνη της, ανακαλύπτοντας πιθανές απειλές και ευκαιρίες.



Εικόνα 4: Μάθηση μέσω εμπειρίας (Πηγή: Bertolini et al., 2021).

3.4.2 Τρόπος Εκπαίδευσης

Offline

Στην offline μάθηση, το στοιχείο της μάθησης δεν συμβαίνει ενεργά κατά τη διάρκεια του gameplay. Η διαδικασία μάθησης λαμβάνει χώρα κυρίως κατά τη φάση ανάπτυξης του παιχνιδιού και ολοκληρώνεται πριν το παιχνίδι κυκλοφορήσει στην αγορά. Για παράδειγμα, μπορεί να γίνει προσπάθεια αντιγραφής του στυλ ή της συμπεριφοράς ενός οδηγού αγώνων ράλι ή ενός ντράμερ με σκοπό τη ρεαλιστική αναπαράσταση μιας αντίστοιχης εικονικής οντότητας (Bertens et al., 2018).

Αυτή η τεχνική είναι παρόμοια με την τεχνική σύλληψης κίνησης (motion capture), η οποία χρησιμοποιείται εδώ και χρόνια στη βιομηχανία των παιχνιδιών, κυρίως σε αθλητικά παιχνίδια. Αυτή η διαδικασία είναι πολύπλοκη καθώς λαμβάνει υπόψη τις χρονικές στιγμές για τις μεταβάσεις από ένα animation σε ένα άλλο.

Ένα παράδειγμα offline μάθησης είναι το παιχνίδι ReVolt (αγώνες αυτοκινήτων), όπου η μηχανή TN εκπαιδεύτηκε στο περιβάλλον χρησιμοποιώντας γενετικούς αλγορίθμους για την εύρεση των βέλτιστων διαδρομών στις πίστες. Αυτός ο τρόπος εκπαίδευσης έχει το πλεονέκτημα ότι το τελικό παιχνίδι μπορεί να υποβληθεί σε συνήθεις διαδικασίες ελέγχου ποιότητας πριν κυκλοφορήσει. Επιπλέον, δεν προσθέτει επιπλέον υπολογιστικό κόστος κατά τη διάρκεια του gameplay, καθώς ο αλγόριθμος μάθησης δεν είναι ενεργός. Ωστόσο, ανάλογα με τον τύπο του αλγορίθμου μάθησης, η αποκτηθείσα γνώση μπορεί να είναι δύσκολο έως αδύνατο να αξιολογηθεί και να εκσφαλματωθεί (Bertens et al., 2018).

Online

Σε αντίθεση με την offline μάθηση, η online μάθηση μπορεί να παραμένει ενεργή και κατά τη διάρκεια του gameplay. Η εκπαίδευση μπορεί να λάβει χώρα εξ ολοκλήρου online, δηλαδή να ξεκινήσει ταυτόχρονα με το gameplay, ή μπορεί να έχει προηγηθεί κάποιας μορφής offline εκπαίδευση. Η online μάθηση επιτρέπει στη μηχανή ΤΝ να προσαρμόζεται στον τρόπο παιχνιδιού κάθε χρήστη και να αναγνωρίζει τις ιδιαιτερότητές του (Bertens et al., 2018).

Για παράδειγμα, στο παιχνίδι Forza Motorsport, επιχειρείται η κατανόηση του στιλ ενός χρήστη με σκοπό τη δημιουργία ενός drivatar (οδηγού που προσομοιώνει το στιλ και τη συμπεριφορά του χρήστη) που θα αγωνίζεται στο παιχνίδι. Ένα άλλο παράδειγμα online μάθησης είναι η δυνατότητα της μηχανής ΤΝ να εκμάθει τα μονοπάτια που χρησιμοποιούν οι ανθρώπινοι χρήστες σε μια εσωτερική έκδοση του παιχνιδιού Command & Conquer: Renegade.

Πρέπει να σημειώσουμε ότι η online μάθηση πρέπει να χρησιμοποιείται με προσοχή, καθώς μπορεί εύκολα να οδηγήσει σε μη επιθυμητές συμπεριφορές. Γι' αυτό το λόγο, συνήθως περιορίζονται οι χώροι καταστάσεων που μπορεί να εξερευνήσει η διαδικασία μάθησης, είτε με φραγή της χρησιμοποιούμενης αναπαράστασης γνώσης είτε με την επιβολή "ενστίκτων" ή κανόνων που η αποκτηθείσα γνώση δεν μπορεί να παραβιάσει. Επιπλέον, θα ήταν καλό να εξεταστεί η δυνατότητα επιλογής ενεργοποίησης/απενεργοποίησης της μάθησης από το χρήστη. Τέλος, πρέπει να σημειωθεί ότι με την online μάθηση υπάρχει πιθανότητα για σημαντικό επιπλέον υπολογιστικό κόστος, ανάλογα με τον τύπο της υλοποίησης του αλγορίθμου μάθησης (Bertens et al., 2018).

3.5 Τεχνικές Μηχανικής Μάθησης

3.5.1 Δένδρα Απόφασης

Τα δένδρα απόφασης χρησιμοποιούνται σε ηλεκτρονικά παιχνίδια εδώ και πάνω από 20 χρόνια ως μέθοδος λήψης αποφάσεων. Με τη χρήση αλγορίθμων μηχανικής μάθησης, όπως το ID3, το C4.5 κ.λπ., είναι δυνατόν να δημιουργηθεί ένα μοντέλο που να μοντελοποιεί τη συμπεριφορά της μηχανής ΤΝ βάσει κάποιων χαρακτηριστικών. Για να γίνει αυτό, απαιτείται η ύπαρξη παραδειγμάτων εκπαίδευσης, τα οποία θα μπορούσαν να συλλεχθούν από πλατφόρμες διαδικτυακών παιχνιδιών (Mania, Guy & Recht, 2018).

Ένα πολύ σημαντικό χαρακτηριστικό των δένδρων απόφασης είναι ότι παράγουν μοντέλα που είναι εύκολα κατανοητά. Αυτό μπορεί να διευκολύνει τις διαδικασίες βελτιστοποίησης και ελέγχου. Για

παράδειγμα, καθώς διατρέχουμε το δέντρο από τη ρίζα προς τα φύλλα, μπορούμε εύκολα να διακρίνουμε τα χαρακτηριστικά μιας συγκεκριμένης συμπεριφοράς και να σημειώσουμε ποια απόφαση ελήφθη σε κάθε κόμβο. Ένα πρόσφατο παράδειγμα παιχνιδιού που χρησιμοποιεί δένδρα απόφασης είναι το Black & White (Mania, Guy & Recht, 2018).

3.5.2 Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα είναι μία από τις πιο διαδεδομένες τεχνικές Μηχανικής Μάθησης στα ηλεκτρονικά παιχνίδια. Παρόλα αυτά, ο τρόπος λειτουργίας τους δεν είναι πάντα απολύτως κατανοητός, γεγονός που δημιουργεί ανησυχίες σε εκείνους που σκέφτονται να τα χρησιμοποιήσουν. Όπως και στα δένδρα απόφασης, τα νευρωνικά δίκτυα απαιτούν δεδομένα εκπαίδευσης πάνω στα οποία θα βασιστεί η εκπαίδευσή τους (Gülü et al., 2023).

Τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για πολλούς σκοπούς στα παιχνίδια, όπως η ταξινόμηση/προσέγγιση συναρτήσεων, η μάθηση πρόβλεψης της ανταμοιβής σε κάποια κατάσταση, η αποτίμηση κατάστασης / ταξινόμηση και πρόβλεψη των ενεργειών του αντιπάλου. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε νευρωνικά δίκτυα για τον έλεγχο της συμπεριφοράς ενός bot σε ένα First-Person Shooter παιχνίδι.

Στο παραπάνω σενάριο, αρχικά κωδικοποιούμε διάφορες τιμές σχετικές με το bot και το περιβάλλον του ως είσοδοι του δικτύου. Στη συνέχεια, καθορίζουμε δείκτες για τις συμπεριφορές του, όπως "ΦΥΓΕ" ή "ΠΥΡΟΒΟΛΗΣΕ", ως έξοδοι του δικτύου. Καθώς το δίκτυο εκπαιδεύεται, μαθαίνει να αντιδρά κατάλληλα στις διάφορες καταστάσεις και να προσαρμόζει τη συμπεριφορά του σύμφωνα με τις ενέργειες του παίκτη. Ωστόσο, πρέπει να λάβουμε υπόψη ότι πρέπει να προσαρμοστούν πολλοί παράγοντες για να εξασφαλιστεί ότι το τελικό δίκτυο δεν θα υπερπροσαρμοστεί (Gülü et al., 2023).

3.5.3 Γενετικοί Αλγόριθμοι

Παρά την αρχική σκληρή κριτική από μέλη της βιομηχανίας παιχνιδιών, οι γενετικοί αλγόριθμοι έχουν αποδειχθεί ως ισχυρό εργαλείο για την τεχνητή νοημοσύνη στα παιχνίδια. Πολλοί σχεδιαστές παιχνιδιών υποστηρίζουν ότι, αν και οι γενετικοί αλγόριθμοι απαιτούν πολλούς υπολογισμούς και είναι αργοί για να παράγουν αποτελέσματα όταν η μάθηση είναι online, όταν η εκπαίδευση γίνεται offline αυτό το αντεπιχείρημα υπονομεύεται, καθώς το κόστος "πληρώνεται" μια φορά κατά τη φάση της ανάπτυξης και εκπαίδευσης (Summerville et al., 2018).

Έτσι, προσεγγίσεις βασισμένες σε γενετικούς αλγορίθμους μπορούν να παρέχουν την κατάλληλη προσαρμοστικότητα στις συμπεριφορές των χαρακτήρων, οδηγώντας σε αναδυόμενες συμπεριφορές. Οι δυσκολίες που μπορούν να προκύψουν στο "σχεδιασμό" αρκετών στρατηγικών μπορούν να αντιμετωπιστούν με τη χρήση γενετικών αλγορίθμων. Παιχνίδια που έχουν εφαρμόσει με επιτυχία γενετικούς αλγορίθμους περιλαμβάνουν το ReVolt για την αναζήτηση των βέλτιστων αγωνιστικών διαδρομών, τη σειρά Creatures, τα Seaman, Nooks & Crannies και το Return Fire II (Summerville et al., 2018).

3.5.4 Μάθηση κατά Bayes

Η μάθηση βασισμένη στο θεώρημα του Bayes αποτελεί μια ελπιδοφόρα προσέγγιση όσον αφορά την εφαρμογή της σε ηλεκτρονικά παιχνίδια. Ένα χαρακτηριστικό της είναι η ικανότητά της να παρέχει αξιόπιστες αναπαραστάσεις για ένα μεγάλο εύρος σεναρίων μάθησης από τον πραγματικό κόσμο, και συνεπώς μπορεί να εφαρμοστεί εύκολα στους εικονικούς κόσμους των παιχνιδιών. Παρότι δεν έχουν αναφερθεί συγκεκριμένα εμπορικά παιχνίδια που χρησιμοποιούν τέτοιες τεχνικές, η ερευνητική κοινότητα έχει προτείνει τη χρήση των μεθόδων Bayes για την εκμάθηση συμπεριφορών σε χαρακτήρες βιντεοπαιχνιδιών, ειδικά σε bots παιχνιδιού πρώτου προσώπου (FPS bots) (Alam, 2022).

Ειδική αναφορά αξίζουν τα Bayesian δίκτυα, τα οποία αποτελούν συμπαγείς αναπαραστάσεις σε μορφή γράφων για τις σχέσεις μεταξύ τυχαίων μεταβλητών σε ένα πρόβλημα. Η χρησιμότητα αυτών των γράφων βασίζεται στη δυνατότητά τους να υποστηρίξουν τη διαδικασία εξαγωγής συμπερασμάτων ή τη λήψη αποφάσεων σε συνθήκες αβεβαιότητας. Η διαδικασία εξαγωγής συμπερασμάτων βασίζεται στο θεώρημα του Bayes για την υπό συνθήκη πιθανότητα. Τα Bayesian δίκτυα θα μπορούσαν να χρησιμοποιηθούν για τη μοντελοποίηση συγκεκριμένων περιπτώσεων όπου οι NPCs απαιτείται να λαμβάνουν αποφάσεις βάσει αβέβαιης πληροφορίας σχετικά με τον κόσμο του παιχνιδιού. Η συνεχής ενημέρωση των πιθανοτήτων που χρησιμοποιούν τα δίκτυα Bayes καθώς το παιχνίδι εξελίσσεται μπορεί να συμβάλει στη δημιουργία ευφυών συστημάτων τεχνητής νοημοσύνης που μαθαίνουν και προσαρμόζονται. Για παράδειγμα, σε ένα παιχνίδι ρόλων (CRPG), το υποσύστημα τεχνητής νοημοσύνης θα μπορούσε να συγκεντρώνει στατιστικά για τις μάχες μεταξύ ενός συγκεκριμένου είδους πλάσματος και μιας συγκεκριμένης κάστας χαρακτήρων, και στη συνέχεια να υπολογίζει την υπό συνθήκη πιθανότητα θνησιμότητας του πλάσματος ανά κλάση χαρακτήρα, προκειμένου αυτό να αποφασίσει εάν θα συμμετάσχει σε μάχη. Μια ακόμα πιο εξελιγμένη προσέγγιση θα ήταν η δημιουργία ενός συστήματος όπου το πλάσμα θα επιλέγει την καλύτερη διαθέσιμη "κα-

τάσταση" για αναζήτηση χαρακτήρων της κάστας, με τη μικρότερη πιθανότητα θνησιμότητας (Alam, 2022).

3.6 Ενισχυτική Μάθηση

3.6.1 Γενικά

Η ενισχυτική μάθηση αναπτύσσεται ως μια ομάδα τεχνικών που έχουν ως βασικό χαρακτηριστικό την αναζήτηση αποτελεσματικών στρατηγικών για την αντιμετώπιση προβλημάτων μέσω εκτεταμένης αυτομάθησης. Η ενισχυτική μάθηση επιτρέπει στους σχεδιαστές ΤΝ να αποφύγουν την ανάγκη να σχεδιάσουν πολιτικές για την επίτευξη στόχων, καθώς απλοποιεί το πρόβλημα στον καθορισμό των στόχων μέσω του σήματος ανταμοιβής. Αυτό σημαίνει ότι είναι κατάλληλη για την αντιμετώπιση προβλημάτων όπου η ευριστική προσέγγιση είναι δύσκολη. Στην ενισχυτική μάθηση, η εκπαίδευση πραγματοποιείται μέσω δοκιμαστικών προσπαθειών και λαθών, με αποτέλεσμα οι προαπαιτήσεις για προηγούμενη γνώση να είναι ελάχιστες (Yannakakis & Togelius, 2018).

Ένας πράκτορας που χρησιμοποιεί ενισχυτική μάθηση αλληλεπιδρά με το περιβάλλον του και ανταμείβεται με βάση μια συνάρτηση ανταμοιβής για κάθε δράση που εκτελεί. Ο κύριος στόχος του πράκτορα είναι η μεγιστοποίηση αυτής της συνάρτησης, προκειμένου να αυξηθεί η συνολική ανταμοιβή. Μία πρώτη εφαρμογή της ενισχυτικής μάθησης είναι η δημιουργία τεχνητών οντοτήτων ζωής, οι οποίες ξεκινούν χωρίς προηγούμενη γνώση και επιδιώκουν να μάθουν, για παράδειγμα, πώς να επιβιώσουν εξερευνώντας το περιβάλλον τους.

Σε περιβάλλοντα παιχνιδιών, όπου συχνά υπάρχουν πολλαπλοί πράκτορες, η απόδοση μιας στρατηγικής εξαρτάται τόσο από την αντίδραση του περιβάλλοντος όσο και από τις ενέργειες των άλλων πρακτόρων. Στην αρχή της διαδικασίας εκμάθησης, οι ενέργειες των πρακτόρων είναι ακαθόριστες, καθιστώντας αδύνατη την εκτέλεση των διάφορων σεναρίων εκπαίδευσης. Μία προσέγγιση που μπορεί να χρησιμοποιηθεί είναι να τοποθετηθούν όλοι οι πράκτορες στο περιβάλλον και να εκπαιδευτούν μέσω αλληλεπίδρασης, ξεκινώντας με τυχαίες ενέργειες και βελτιώνοντας σταδιακά τις ικανότητές τους βάσει των ανταμοιβών που λαμβάνουν. Καθώς όλοι οι πράκτορες μαθαίνουν ταυτόχρονα, πρέπει να αντιμετωπίσουν ισχυρότερους αντιπάλους σε κάθε στάδιο της διαδικασίας, γεγονός που τους επιτρέπει να προσαρμοστούν σε διάφορες στρατηγικές αντιπάλων (Yannakakis & Togelius, 2018).

3.6.2 Δυναμική Παραγωγή Σεναρίων

Η δυναμική παραγωγή σεναρίων (dynamic scripting) είναι μια τεχνική μηχανικής μη επιβλεπόμενης μάθησης για ηλεκτρονικά παιχνίδια, η οποία μπορεί να χαρακτηριστεί ως στοχαστική βελτιστοποίηση, επηρεασμένη από την αρχιτεκτονική κριτή-δράστη. Αυτή η τεχνική βασίζεται σε πολλές βάσεις κανόνων, με μία για κάθε κλάση πράκτορα στο παιχνίδι. Κάθε φορά που δημιουργείται μια νέα «περίσταση» πράκτορα, οι βάσεις αυτές χρησιμοποιούνται για να δημιουργηθεί ένα νέο σενάριο που καθορίζει τη συμπεριφορά του. Τα σενάρια εξάγονται από τις βάσεις κανόνων σύμφωνα με την κλάση του πράκτορα και ελέγχουν τον πράκτορα σε συγκεκριμένες καταστάσεις (Frutos-Pascual & Zapirain, 2015).

Ο στόχος της τεχνικής είναι η προσαρμογή των βαρών στις βάσεις κανόνων, ώστε η αναμενόμενη «υγεία» της συμπεριφοράς που ορίζεται από τα παραγόμενα σενάρια να αυξάνεται με γρήγορους ρυθμούς, ακόμα και σε πολύ μεταβλητά περιβάλλοντα. Οι κανόνες αξιολογούνται με βάση τη συνεισφορά τους στο τελικό αποτέλεσμα, με τους επιτυχημένους κανόνες να επιβραβεύονται και τους ανεπιτυχής να τιμωρούνται. Η προσαρμογή των βαρών κάθε κανόνα γίνεται με ανταλλαγή βαρών μεταξύ των κανόνων, διατηρώντας το συνολικό άθροισμα σταθερό.

Αυτή η τεχνική έχει χρησιμοποιηθεί με επιτυχία στο Neverwinter Nights μέσω ενός σχετικού module που έχει αναπτυχθεί (Frutos-Pascual & Zapirain, 2015).

3.7 Προοπτικές

3.7.1 Γενικές

Η εφαρμογή της Μηχανικής Μάθησης μπορεί να αποδειχτεί επωφελής σε σχέση με τις ανάγκες υπολογιστικής ισχύος, καθώς η χρήση της μπορεί να αντικαταστήσει τον επαναλαμβανόμενο χρονοπρογραμματισμό με την αξιοποίηση συσσωρευμένης γνώσης. Ωστόσο, υπάρχει το ενδεχόμενο να συμβεί το ακριβώς αντίθετο, καθώς ο αλγόριθμος μάθησης μπορεί να επιβάλλει επιπρόσθετο φόρτο στη χρήση CPU και μνήμη (Shao et al., 2019).

3.7.2 Προοπτικές για τη Βιομηχανία

Από την πλευρά της βιομηχανίας ανάπτυξης ηλεκτρονικών παιχνιδιών, η εφαρμογή προηγμένων τε-

χνικών Τεχνητής Νοημοσύνης (ΤΝ), ειδικά η Μηχανική Μάθηση, μπορεί να οδηγήσει σε μειωμένο κόστος ανάπτυξης του υποσυστήματος ΤΝ, καθώς αποφεύγεται ο χρονοβόρος και επίπονος "χειροκίνητος" προγραμματισμός συμπεριφορών. Επιπλέον, κατά την προώθηση των παιχνιδιών, η χρήση τεχνικών Μηχανικής Μάθησης μπορεί να αναδειχθεί ως σημαντικό πλεονέκτημα, δημιουργώντας μάρκετινγκ ενδιαφέρον και προσελκύοντας το ενδιαφέρον του αγοραστικού κοινού. Όλα αυτά μπορούν να συντελέσουν στο να υπερνικήσει η βιομηχανία παιχνιδιών την προκατάληψη όσον αφορά τη χρήση τεχνικών Μηχανικής Μάθησης και άλλων προηγμένων τεχνικών ΤΝ (Shao et al., 2019).

Από την άλλη πλευρά, η χρήση τεχνικών Μηχανικής Μάθησης ενδέχεται να οδηγήσει σε αύξηση του κόστους ανάπτυξης των υποσυστημάτων Τεχνητής Νοημοσύνης. Αφενός, η έλλειψη προγραμματιστών με εμπειρία στη Μηχανική Μάθηση μπορεί να αποτελέσει πρόκληση. Αφετέρου, αυξάνεται ο χρόνος που απαιτείται για την ανάπτυξη, τον έλεγχο και την αντιμετώπιση σφαλμάτων του αλγορίθμου μάθησης. Ιδιαίτερα κατά τον έλεγχο και την αντιμετώπιση σφαλμάτων, το εύρος των πιθανών συμπεριφορών που πρέπει να ληφθούν υπόψη μπορεί να είναι πολύ μεγαλύτερο όταν χρησιμοποιούνται τεχνικές Μηχανικής Μάθησης. Γενικά, ο έλεγχος ποιότητας αποτελεί ένα από τα πιο κρίσιμα και δύσκολα σημεία όσον αφορά τη χρήση τεχνικών Μηχανικής Μάθησης σε παιχνίδια (Shao et al., 2019).

3.7.3 Επιδράσεις στο Gameplay

Η ποικιλία των εικονικών κόσμων στα ηλεκτρονικά παιχνίδια και, ως εκ τούτου, η ποικιλία των προκλήσεων που σχετίζονται με τη Μηχανική Μάθηση, ουσιαστικά περιορίζεται μόνο από τη φαντασία. Η εφαρμογή τεχνικών Μηχανικής Μάθησης θα μπορούσε να βελτιώσει τη συμπεριφορά των τεχνητών ευφυών οντοτήτων και να οδηγήσει σε βελτιστοποίηση των κανόνων, του περιβάλλοντος, των υποδομών και των διεπαφών του παιχνιδιού. Η συμπεριφορά των διαφόρων οντοτήτων μπορεί να γίνει πιο δυναμική, πιο πιστευτή, πιο προκλητική και πιο εύρωστη. Επιπλέον, η χρήση τεχνικών Μηχανικής Μάθησης μπορεί να δημιουργήσει μια πιο προσαρμοσμένη εμπειρία παιχνιδιού και να αυξήσει την επαναληψιμότητα του παιχνιδιού. Αυτό μπορεί να συμβεί μέσω του γεγονότος ότι η ΤΝ εξελίσσεται καθώς εξελίσσεται και ο παίκτης, μαθαίνοντας τα δεδομένα σχετικά με τις προτιμήσεις και τις στρατηγικές του (de Almeida Rocha & Duarte, 2019).

Πέραν των θετικών πτυχών που προκύπτουν από τη χρήση της Μηχανικής Μάθησης στα παιχνίδια, υπάρχει η πιθανότητα να προκύψουν συμπεριφορές που μπορεί να φανούν "περίεργες" στον παίκτη, καθώς αυτές δεν είναι πλήρως ελεγχόμενες από τους σχεδιαστές. Επιπλέον, μπορεί να είναι δύ-

σκολο για τον παίκτη να προβλέψει όλες τις μελλοντικές συμπεριφορές που μπορεί να προκύψουν κατά τη διάρκεια της εξέλιξης της μάθησης, δημιουργώντας ανασφάλεια. Επιπλέον, υπάρχει η πιθανότητα η μάθηση να "κολλήσει" σε ένα σημείο όπου η απόδοση του αλγορίθμου μάθησης δεν είναι ικανοποιητική, χωρίς δυνατότητα διόρθωσης. Τέλος, σε περίπτωση που η διαδικασία μάθησης είναι online, υπάρχει πάντα η πιθανότητα η μηχανή ΤΝ να μάθει από παραδείγματα που πρέπει να αποφεύγονται, κάτι που θα οδηγήσει και αυτή σε ανεπιθύμητη συμπεριφορά (de Almeida Rocha & Duarte, 2019).

3.8 Άλλες Σχετικές Εφαρμογές

Πέραν των εφαρμογών που άπτονται άμεσα του gameplay, υπάρχουν και άλλες σχετικές εφαρμογές που συνδέονται με τα παιχνίδια και εξετάζουν διάφορα ζητήματα. Μια ενδιαφέρουσα προοπτική αποτελεί η ανάπτυξη παιχνιδιών βασισμένη σε δεδομένα (Data-driven Game Development), όπου τα πραγματικά δεδομένα μπορούν να χρησιμοποιηθούν για τη δημιουργία ρεαλιστικών γραφικών, περιβαλλόντων και συνθηκών εξομοίωσης στο εικονικό περιβάλλον. Πολλές πιθανές εφαρμογές αυτής της προσέγγισης αφορούν τα παιχνίδια πολλαπλών παικτών, τα οποία μπορούν επίσης να χρησιμοποιηθούν για τη συλλογή δεδομένων εκπαίδευσης (Bertolini et al., 2021).

Με τη χρήση της μηχανικής μάθησης, μπορεί να επιτευχθεί βελτιωμένος αλγόριθμος για την κατάταξη των παικτών σε παιχνίδια πολλαπλών παικτών και για την ανίχνευση τυχόν gamebots. Επιπλέον, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την ανάλυση του gameplay και τον έλεγχο ποιότητας των παιχνιδιών. Τέλος, στο πλαίσιο των λεγόμενων enserious games, παιχνίδια με έντονο εκπαιδευτικό προσανατολισμό, η χρήση τεχνικών Μηχανικής Μάθησης ανοίγει ενδιαφέρουσες προοπτικές που αξίζει να εξερευνηθούν (Bertolini et al., 2021).

4 Αλγόριθμος Ενισχυτικής Μάθησης σε περιβάλλον παιχνιδιού

Για την εκπαίδευση ενός πράκτορα βαθιάς ενισχυτικής μάθησης, απαιτείται η ανάπτυξη προγράμματος για τον αλγόριθμο της εκπαίδευσης και την αλληλεπίδραση του πράκτορα με το περιβάλλον του. Στην συνέχεια περιγράφεται το λογισμικό που χρησιμοποιήθηκε για την ανάπτυξη του προγράμματος.

4.1 Python

Η απόφαση να χρησιμοποιηθεί η Python ως γλώσσα προγραμματισμού βασίστηκε στο ποικίλο φάσμα των δυνατοτήτων της. Η Python είναι γνωστή ως μια διερμηνευόμενη, γενικού σκοπού και υψηλού επιπέδου γλώσσα προγραμματισμού. Ακολουθεί μια προγραμματιστική ιδεολογία που τονίζει τη σημασία του καθαρού, απλού και κατανοητού κώδικα, αρχές που περιγράφονται στο μανιφέστο «The Zen of Python».

Από τεχνική άποψη, η Python έχει το πλεονέκτημα ότι μπορεί να υποστηρίζει τόσο διαδικαστικά όσο και αντικειμενοστραφή στυλ προγραμματισμού. Αυτή η ευελιξία επιτρέπει στους προγραμματιστές να έχουν μεγαλύτερη ευελιξία κατά το σχεδιασμό και την υλοποίηση των προγραμμάτων τους. Ένα από τα βασικά χαρακτηριστικά της Python είναι η δυναμική πληκτρολόγηση, που σημαίνει ότι οι προγραμματιστές δεν χρειάζεται να καθορίζουν τύπους για μεταβλητές, καθιστώντας τον κώδικα πιο συνοπτικό και ευκολότερο να γραφτεί. Ένα άλλο πλεονέκτημα είναι η λειτουργία αυτόματης συλλογής σκουπιδιών, η οποία φροντίζει για τη διαχείριση της κατανομής μνήμης, μειώνοντας τις πιθανότητες διαρροής μνήμης και καθιστώντας τη διαδικασία προγραμματισμού πιο αποτελεσματική. Επιπλέον, η ερμηνευτική φύση της Python της επιτρέπει να εκτελείται σε διαφορετικές πλατφόρμες χωρίς να χρειάζεται καμία τροποποίηση. Αυτή η συμβατότητα μεταξύ πλατφορμών είναι ένα τεράστιο πλεονέκτημα, καθώς εξοικονομεί χρόνο και προσπάθεια για προγραμματιστές που θέλουν ο κώδικάς τους να εκτελείται απρόσκοπτα σε διάφορα λειτουργικά συστήματα. Συνολικά, η υποστήριξη της Python τόσο για διαδικαστικά όσο και για αντικειμενοστραφή παραδείγματα προγραμματισμού, η δυναμική πληκτρολόγηση, η συλλογή σκουπιδιών, η ανεξαρτησία της πλατφόρμας και το πλούσιο οικοσύστημα βιβλιοθηκών συμβάλλουν στο να γίνει μια απίστευτα ευέλικτη και πρακτική γλώσσα για να δουλέψουν οι προγραμματιστές. Επιπλέον, η Python διαθέτει μια εκτενή συλλογή βιβλιοθηκών που καλύπτουν ένα ευρύ φάσμα λειτουργιών. Αυτές οι βιβλιοθήκες παρέχουν έτοιμες λύσεις τόσο για εξειδικευμένες όσο και για γενικές εργασίες, πράγμα που σημαίνει ότι οι προγραμματιστές δεν χρειάζεται να ανακαλύπτουν ξανά τον τροχό κάθε φορά που αντιμετωπίζουν ένα κοινό πρόβλημα προγραμματισμού. Αυτό το τεράστιο οικοσύστημα βιβλιοθήκης ενισχύ-

ει σημαντικά την πρακτικότητα της Python και την καθιστά μια εξαιρετικά κατάλληλη γλώσσα για πολλές εφαρμογές.

4.1.1 Βιβλιοθήκες

Στην εργασία έγινε εκτεταμένη χρήση βιβλιοθηκών τόσο για την επεξεργασία των δεδομένων, όσο και για την κατασκευή και εκπαίδευση των νευρονικών δικτύων. Οι βιβλιοθήκες που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

Gym

Η πλατφόρμα Gym, που αναπτύχθηκε από την OpenAI, κερδίζει την αναγνώριση ως κορυφαίο εργαλείο για την ανάπτυξη και την αξιολόγηση αλγορίθμων ενισχυτικής μάθησης. Αυτή η βιβλιοθήκη διαθέτει μια εντυπωσιακή σειρά περιβαλλόντων, παρέχοντας στους ερευνητές την ελευθερία να εφαρμόσουν και να αξιολογήσουν τους αλγόριθμούς τους χωρίς κανέναν περιορισμό στο σχεδιασμό τους. Επιπλέον, η συμβατότητά του με δημοφιλείς υπολογιστικές βιβλιοθήκες όπως οι PyTorch, TensorFlow και Theano επιτρέπει τεράστιες δυνατότητες όσον αφορά τις εφαρμογές.

Ο λόγος για τον οποίο υπήρχε ζήτηση για ένα εργαλείο όπως το Gym ήταν επειδή υπήρχε έλλειψη σημείων αναφοράς που θα μπορούσαν να χρησιμοποιηθούν για τη σύγκριση διαφορετικών αλγορίθμων ενισχυτικής μάθησης και την αξιολόγηση της αξιοπιστίας των σχετικών ερευνητικών εργασιών. Προκειμένου να ανταποκριθεί σε αυτήν την ανάγκη, το Gym προσφέρει ένα ευρύ φάσμα περιβαλλόντων που είναι βολικά οργανωμένα σε ξεχωριστές κατηγορίες.

Η βιβλιοθήκη Gym προσφέρει μια ποικιλία κατηγοριών για τη δοκιμή αλγορίθμων, συμπεριλαμβανομένων των Classic Control και Toy-text, που εστιάζουν σε απλά προβλήματα μικρής κλίμακας. Η κατηγορία Algorithmic προκαλεί τους πράκτορες με αριθμητικούς υπολογισμούς. Η κατηγορία Atari παρέχει περιβάλλοντα που βασίζονται στην πλατφόρμα παιχνιδιών Atari, επιτρέποντας τη δοκιμή αλγορίθμων σε ένα περιβάλλον που μοιάζει με παιχνίδι. Η κατηγορία 2D-3D Robots επιτρέπει την προσομοίωση ρομποτικών συστημάτων χρησιμοποιώντας τη μηχανή φυσικής MuJoCo. Το Gym περιλαμβάνει επίσης ειδικά πακέτα διαμόρφωσης για περιβάλλοντα όπως το Atari για να κάνουν τους εκπαιδευτικούς πράκτορες πιο αποτελεσματικούς.

Numpy

Η NumPy είναι μια πολύ σημαντική βιβλιοθήκη στη Python, ειδικά στον τομέα των επιστημονικών υπολογιστών. Είναι ιδιαίτερα γνωστή για την ικανότητά της να χειρίζεται αποτελεσματικά μεγάλους και πολυδιάστατους πίνακες αξιοποιώντας διανυσματικούς υπολογισμούς. Επιπλέον, η NumPy προσφέρει ένα εκτεταμένο εύρος προηγμένων μαθηματικών συναρτήσεων που απλοποιούν σημαντικά την εκτέλεση σύνθετων υπολογιστικών πράξεων σε πίνακες με μεγάλο αριθμό διαστάσεων.

Ένα από τα κύρια πλεονεκτήματα της NumPy είναι η ικανότητά της να εκτελεί υπολογισμούς με απίστευτα γρήγορο ρυθμό μέσω της χρήσης διανυσματοποίησης. Αυτή η προηγμένη δυνατότητα επιτρέπει στους χρήστες να γράφουν μαθηματικές εκφράσεις που μπορούν να επεξεργαστούν ολόκληρους πίνακες με μία μόνο εντολή, εξαλείφοντας την ανάγκη για επαναλαμβανόμενες επαναλήψεις. Ως αποτέλεσμα, αυξάνεται η ταχύτητα των υπολογισμών και βελτιώνεται η αποδοτικότητα των προγραμμάτων.

Η PyTorch, αναπτυγμένη από την ερευνητική ομάδα τεχνητής νοημοσύνης του Facebook (FAIR), αθεωρείται μία από τις κορυφαίες βιβλιοθήκες ανοιχτού κώδικα για βαθιά μηχανική μάθηση, παράλληλα με την TensorFlow.

PyTorch

Η βιβλιοθήκη αυτή σχεδιάστηκε με σκοπό την εκμετάλλευση της δύναμης των καρτών γραφικών (GPU), προσφέροντας σημαντική ταχύτητα και ευελιξία στους χρήστες.

Ένα βασικό χαρακτηριστικό της PyTorch είναι οι Tensors, οι οποίοι είναι πίνακες δομής παρόμοιας με αυτούς της NumPy, αλλά με επιπλέον δυνατότητες προσαρμοσμένες για υψηλές απαιτήσεις υπολογιστικών επιδόσεων. Οι Tensors διατίθενται σε δύο εκδόσεις ανάλογα με την υπολογιστική μο-

νάδα (CPU ή GPU) που θα χρησιμοποιήσει ο χρήστης, και επιτρέπουν την αυτόματη διαχείριση του υπολογισμού κλίσεων (gradients), έναν κρίσιμο παράγοντα στην εκπαίδευση νευρωνικών δικτύων.

Η αυτόματη διαφόριση (automatic differentiation) είναι το δεύτερο σημαντικό χαρακτηριστικό της PyTorch, που επιτρέπει την εύκολη και αποτελεσματική κατασκευή πολύπλοκων βαθιών νευρωνικών δικτύων. Αυτή η λειτουργία υποστηρίζει τον υπολογισμό της κλίσης σε επίπεδα των δικτύων, διευκολύνοντας την εφαρμογή μεθόδων βελτιστοποίησης και ανανέωσης των βαρών του δικτύου κατά τη διάρκεια της εκπαίδευσης. Η χρήση GPU επιταχύνει αυτούς τους υπολογισμούς, το οποίο οδηγεί σε βελτιώσεις στις επιδόσεις και στην ταχύτητα επεξεργασίας των δεδομένων.

TensorBoard

Η TensorBoard αποτελεί ένα εργαλείο απεικόνισης που συνδέεται με τη βιβλιοθήκη TensorFlow. Αυτό το πακέτο επιτρέπει στους χρήστες να καταγράφουν και να οπτικοποιούν διάφορες τιμές, όπως μετρικές απόδοσης, παραμέτρους, και καταστάσεις εκπαίδευσης νευρωνικών δικτύων μέσω γραφημάτων και άλλων οπτικών απεικονίσεων.

Στο πεδίο της μηχανικής μάθησης, η ικανότητα καταγραφής και ανάλυσης διάφορων μετρικών είναι κρίσιμης σημασίας, καθώς παρέχει στους ερευνητές σαφείς ενδείξεις για την πρόοδο και την απόδοση των μοντέλων τους κατά τη διάρκεια της εκπαίδευσης. Η TensorBoard επιτρέπει την ταυτόχρονη παρουσίαση δεδομένων από πολλαπλές προπονήσεις, διευκολύνοντας έτσι τη σύγκριση μεταξύ διαφορετικών μοντέλων ή ρυθμίσεων. Χάρη σε αυτήν την ολοκληρωμένη προβολή, οι χρήστες μπορούν να εξάγουν συμπεράσματα σχετικά με το πώς οι αλλαγές σε υπερπαραμέτρους ή τοπολογίες επηρεάζουν την απόδοση του συστήματος.

GoogleColab

Το Google Colab, γνωστό απλώς ως Colab, είναι ένα δωρεάν διαδικτυακό εργαλείο που λειτουργεί εξολοκλήρου στο cloud, επιτρέποντας την εκτέλεση, δημιουργία, και κοινή χρήση Jupyter notebooks απευθείας μέσω Google Drive. Αναπτύχθηκε από το τμήμα έρευνας της Google και συνδέεται άμεσα με τον Google λογαριασμό του χρήστη, προσφέροντας πρόσβαση σε διάφορα είδη virtual machines (VM).

Χρήστες του Colab μπορούν να γράφουν κώδικα σε Python, να χειρίζονται σημειωματάρια Jupyter με ευκολία και να ανταλλάσσουν δεδομένα μέσω GitHub ή Google Drive, προσθέτοντας επιπλέον άνεση και ευελιξία στη διαχείριση έργων. Επίσης, επιτρέπει την εισαγωγή και χρήση δεδομένων από πλατφόρμες όπως Kaggle, καθώς και την εγκατάσταση και χρήση προεγκατεστημένων ή τρίτων βιβλιοθηκών Python, όπως PyTorch, TensorFlow, Keras και OpenCV.

Ένα από τα μεγάλα πλεονεκτήματα του Colab είναι η δυνατότητα χρήσης ενισχυμένων υπολογιστικών μονάδων όπως GPU και TPU. Αυτή η δυνατότητα είναι ιδιαίτερα σημαντική για την επιτάχυνση αλγορίθμων μηχανικής μάθησης, ειδικά στην εκπαίδευση βαθιών νευρωνικών δικτύων. Ωστόσο, οι χρήστες πρέπει να λαμβάνουν υπόψη το χρονικό περιθώριο των 12 ωρών για την δέσμευση ενός virtual machine, μετά το οποίο το σημειωματάριο αποσυνδέεται αυτόματα.

4.2 Περιβάλλον

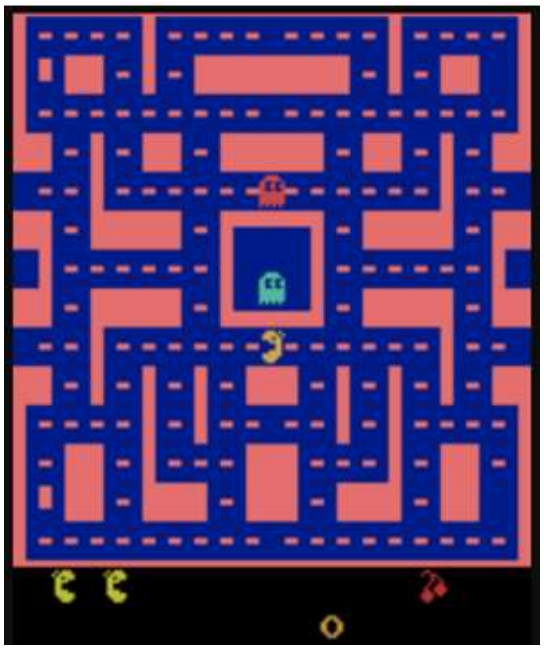
Η πλατφόρμα Arcade Learning Environment (ALE) αποτελεί έναν σημαντικό εργαλείο για την αξιολόγηση αλγορίθμων τεχνητής νοημοσύνης, με έμφαση στην ενισχυτική μάθηση. Αναπτύχθηκε για να παρέχει ένα περιβάλλον όπου πράκτορες μπορούν να εκπαιδεύονται και να δοκιμάζονται μέσα από την αλληλεπίδρασή τους με παιχνίδια της κονσόλας Atari 2600. Τα παιχνίδια αυτά προσφέρουν μια ποικιλία από εργασίες, αποτελώντας ένα δοκιμαστικό πεδίο για την αξιολόγηση της ευελιξίας και της προσαρμοστικότητας των αλγορίθμων.

Ένας από τους πλέον γνωστούς αλγορίθμους που έχει δοκιμαστεί στην ALE είναι ο Deep Q-Network (DQN). Οι πράκτορες στην ALE εκπαιδεύονται χωρίς προηγούμενη γνώση των κανόνων ή της δομής του παιχνιδιού, βασιζόμενοι αποκλειστικά στην αλληλεπίδραση με το περιβάλλον και την ανταμοιβή από αυτή.

Το περιβάλλον παιχνιδιού για την MsPacman θα υλοποιηθεί στην πλατφόρμα ALE(Arcade Learning Environment). Αυτό το παιχνίδι είναι μια τροποποιημένη έκδοση του κλασικού παιχνιδιού Pacman όπου ένας χαρακτήρας παγιδεύεται σε έναν λαβύρινθο, προσπαθώντας να αποφύγει τέσσερα φαντάσματα. Ο χαρακτήρας έχει 9 επιτρεπόμενες κινήσεις:

- Πάνω: Ο χαρακτήρας κινείται προς τα πάνω.
- Κάτω: Ο χαρακτήρας κινείται προς τα κάτω.
- Αριστερά: Ο χαρακτήρας κινείται προς τα αριστερά.
- Δεξιά: Ο χαρακτήρας κινείται προς τα δεξιά.
- Πάνω-Αριστερά: Ο χαρακτήρας κινείται διαγώνια προς τα πάνω και αριστερά.
- Πάνω-Δεξιά: Ο χαρακτήρας κινείται διαγώνια προς τα πάνω και δεξιά.
- Κάτω-Αριστερά: Ο χαρακτήρας κινείται διαγώνια προς τα κάτω και αριστερά.
- Κάτω-Δεξιά: Ο χαρακτήρας κινείται διαγώνια προς τα κάτω και δεξιά.
- Καμία κίνηση: Ο χαρακτήρας παραμένει ακίνητος.

Ο κύριος στόχος του παιχνιδιού είναι ο χαρακτήρας να κερδίσει πόντους συλλέγοντας όσο το δυνατόν περισσότερες μπάλες. Μερικές από αυτές τις μπάλες μετατρέπουν τα φαντάσματα μπλε, επιτρέποντας στον χαρακτήρα να τα εξουδετερώσει και να κερδίσει επιπλέον πόντους. Ο χαρακτήρας ξεκινά με τέσσερις ζωές και χάνει μία κάθε φορά που έρχονται σε επαφή με ένα φάντασμα. Η εξάντληση των ζωών έχει ως αποτέλεσμα το τέλος του παιχνιδιού.



Εικόνα 5 Ένα καρέ του παιχνιδιού MsPacman

Δημιουργία περιβάλλοντος

Η βιβλιοθήκη Gym της OpenAI παρέχει 49 παιχνίδια Atari, το καθένα από τα οποία προσφέρεται σε 12 διαφορετικές εκδόσεις. Αυτές οι εκδόσεις διακρίνονται κυρίως με βάση το είδος των παρατηρήσεων που παρέχονται στους πράκτορες.

Υπάρχουν δύο βασικές κατηγορίες εκδόσεων ανάλογα με τον τύπο των δεδομένων εισόδου:

Εκδόσεις RAM: Σε αυτή την περίπτωση, ο πράκτορας λαμβάνει τις πληροφορίες του παιχνιδιού μέσω ενός διανύσματος 128 bytes RAM, παρέχοντας μια πιο άμεση αντίληψη της εσωτερικής κατάστασης του παιχνιδιού χωρίς οπτική αναπαράσταση.

Εκδόσεις RGB: Σε αυτή την περίπτωση, ο πράκτορας λαμβάνει εικόνες απεικονίζοντας το οπτικό περιβάλλον του παιχνιδιού. Οι εικόνες αυτές είναι τρισδιάστατοι πίνακες διαστάσεων (210, 160, 3), όπου οι τιμές κάθε pixel κυμαίνονται μεταξύ 0 και 255, παρέχοντας μια πλήρη οπτική αναπαράσταση του παιχνιδιού.

Επιπλέον, κάθε μία από αυτές τις κύριες κατηγορίες διαθέτει 6 επιμέρους εκδόσεις. Οι διαφοροποιήσεις μεταξύ των εκδόσεων αυτών οφείλονται κυρίως στην παρουσία ή όχι τυχαιότητας στο περιβάλλον του παιχνιδιού. Μια από τις πιο σημαντικές παραμέτρους που διαφοροποιούν τις εκδόσεις είναι η εφαρμογή των λεγόμενων "sticky actions". Αυτό σημαίνει ότι η δράση του πράκτορα επαναλαμβάνεται για κάποιο αριθμό καρέ, και υπάρχει μια πιθανότητα σ% να επαναληφθεί η προηγούμενη δράση αγνοώντας τη νέα εντολή. Αυτός ο μηχανισμός προσομοιώνει μια μορφή τυχαιότητας και αυξάνει την πρόκληση για τον αλγόριθμο, δοκιμάζοντας την ικανότητά του να ανταποκρίνεται σε απρόβλεπτες συνθήκες.

Όνομα του παιχνιδιού(έκδοση)	**Αριθμός καρέ επανάληψης δράσης**	**Πιθανότητα σ(%)**
-v0	2 ή 3 ή 4	25
-v4	2 ή 3 ή 4	0
Deterministic-v0	4	25
Deterministic-v4	4	0
NoFrameskip-v0	0	25
NoFrameskip-v4	0	0

Πίνακας 1. Ιδιότητες των εκδόσεων των παιχνιδιών της πλατφόρμας Atari.

Επεισοδιακή ζωή

Η ενότητα αυτή συζητά την έννοια της "επεισοδιακής ζωής" σε πλαίσια ενισχυτικής μάθησης, κυρίως στο πλαίσιο παιχνιδιών όπου ο πράκτορας έχει πολλαπλές προσπάθειες ή "ζωές" για να επιτύχει τον στόχο του. Σε αυτό το πλαίσιο, η έλλειψη οποιασδήποτε ποινής για την απώλεια μιας ζωής μπορεί να καθιστά πιο περίπλοκη την εκμάθηση, καθώς ο πράκτορας δεν έχει άμεση αρνητική ανατροφοδότηση. Παρ' όλα αυτά, αυτό τον προτρέπει να μάθει πώς να αποφεύγει τις καταστάσεις που οδηγούν σε απώλεια ζωής, βελτιώνοντας την προσαρμοστικότητά του και τελικά αναπτύσσοντας μια πιο ισχυρή συνάρτηση αξίας. Κάθε "επεισόδιο" στη ζωή του πράκτορα ξεκινά από ένα αρχικό σημείο και καταλήγει στο τερματικό σημείο της απώλειας μιας ζωής. Η διαδικασία αυτή επαναλαμβάνεται μέχρι ο πράκτορας να εξαντλήσει όλες τις διαθέσιμες "ζωές", κατά την οποία το περιβάλλον επανεκκινείται.

Εκκίνηση περιβάλλοντος (FireResetEnv και No-op actions)

Η διαδικασία προετοιμασίας του περιβάλλοντος σε ορισμένα παιχνίδια, όπως το MsPacman, απαιτεί την εκτέλεση μιας ενέργειας πριν από την εκκίνηση του παιχνιδιού.

Για να ξεπεραστεί αυτό το ζήτημα, προβλέπεται η εκτέλεση μιας τυχαίας ενέργειας που ενεργοποιεί την αρχή του παιχνιδιού. Επιπλέον, είναι σύνηθες για τον πράκτορα να ξεκινά πάντα από την ίδια αρχική κατάσταση και να λαμβάνει την ίδια παρατήρηση, κάτι που θα μπορούσε να οδηγήσει στον πράκτορα να απομνημονεύσει αυτήν την κατάσταση και να μην προσαρμοστεί καλά όταν ξεκινάει από διαφορετική θέση. Για να αντιμετωπιστεί αυτό, ο πράκτορας ξεκινά σε μια τυχαία τοποθεσία στην αρχή κάθε επεισοδίου. Η μέθοδος No-op, η οποία περιλαμβάνει την εκτέλεση τυχαίων δράσεων για έναν προκαθορισμένο αριθμό βημάτων στην αρχή κάθε επεισοδίου, συνήθως έως και 30, βοηθά στην ενίσχυση της εκπαίδευσης. Μετά από αυτήν την αρχική περίοδο, ο πράκτορας αρχίζει να λαμβάνει παρατηρήσεις από το περιβάλλον ξεκινώντας από την κατάσταση που έφτασε κατά το τελευταίο βήμα.

Η πολιτική αυτή στοχεύει στην προώθηση μιας πιο ρεαλιστικής απόδοσης από τον πράκτορα, καθώς οδηγεί σε μια πιο ετερογενή αρχική κατάσταση, εκπαιδεύοντάς τον να αντιμετωπίζει μια ευρύτερη ποικιλία σεναρίων και διαφορετικών αρχικών συνθηκών.

Clip reward

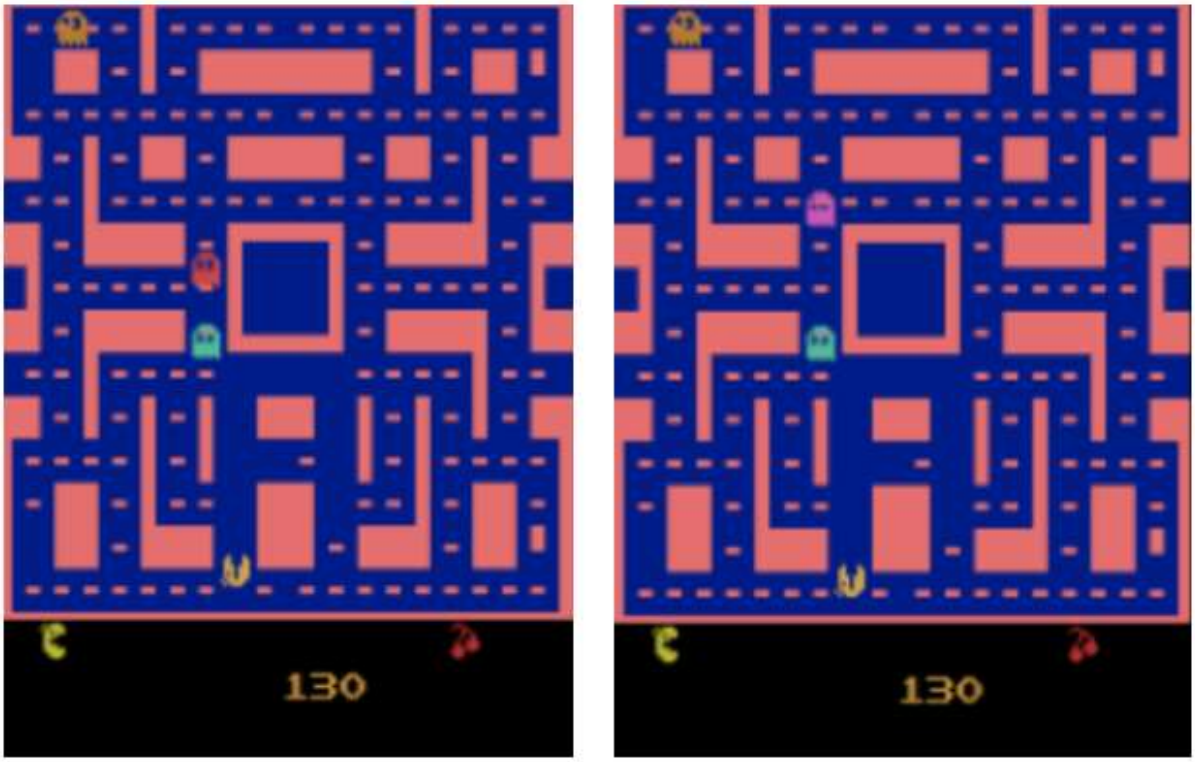
Στο πλαίσιο της ενισχυτικής μάθησης, διαφορετικά περιβάλλοντα ενδέχεται να παρέχουν στους πράκτορες ανταμοιβές με έντονα διαφορετικές κλίμακες τιμών. Για παράδειγμα, στο παιχνίδι Pong, οι ανταμοιβές είναι κανονικοποιημένες στο εύρος από -1 έως 1, ενώ στο MsPacman οι ανταμοιβές μπορούν να φτάσουν μέχρι τους +1600 για την εξουδετέρωση ενός φαντάσματος, και +10 για την κατάποση ενός σφαιριδίου. Αυτή η διακύμανση στο μέγεθος των ανταμοιβών μπορεί να οδηγήσει σε στρατηγικές που εστιάζουν στην απόκτηση της μέγιστης άμεσης ανταμοιβής αντί για την μεγιστοποίηση του συνολικού κέρδους.

Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι το "ψαλίδισμα" των ανταμοιβών, όπου οι ανταμοιβές περιορίζονται σε ένα σταθερό εύρος, συνήθως [-1,1]. Οποιαδήποτε ανταμοιβή που υπερβαίνει αυτό το εύρος μετατρέπεται απλώς σε -1 ή +1, ανάλογα με το πρόσημό της. Αυτό βοηθά στην αποφυγή στρατηγικών που επιδιώκουν την άμεση ανταμοιβή και ενθαρρύνει την ανάπτυξη πιο ισορροπημένων τακτικών αντίδρασης. Εναλλακτικές μέθοδοι μπορεί να εστιάζουν στην καλύτερη εκμετάλλευση του πλήρους φάσματος των ανταμοιβών.

Μέγιστη τιμή εικονοστοιχείου

Σε πλατφόρμες όπως η Arcade Learning Environment (ALE), οι τεχνολογικοί περιορισμοί της μηχανής που φιλοξενούσε την ALE, μπορούν να εισάγουν μη-τυπικές προκλήσεις στη διαδικασία μάθησης των πρακτόρων. Ένα τέτοιο παράδειγμα είναι το τρεμόπαιγμα των εικονοστοιχείων, όπου το ίδιο εικονοστοιχείο παρουσιάζει διαφορετικές τιμές σε διαδοχικές χρονικές στιγμές. Αυτό το φαινόμενο μπορεί να δυσκολέψει τον πράκτορα να αντιληφθεί το πραγματικό περιβάλλον, επηρεάζοντας την απόδοση της μάθησης.

Για να αντιμετωπιστεί αυτό το πρόβλημα, υιοθετήθηκε μια προσέγγιση όπου η τιμή ενός εικονοστοιχείου σε ένα δοσμένο καρέ ορίζεται ως η μέγιστη τιμή μεταξύ της τρέχουσας τιμής και της τιμής του στο προηγούμενο καρέ. Αυτός ο κανόνας εξασφαλίζει ότι το περιβάλλον που αντιλαμβάνεται ο πράκτορας είναι πιο σταθερό. Με αυτό τον τρόπο, αντικείμενα που ίσως εμφανίζονται μόνο σε περιττά ή άρτια καρέ λόγω τρεμοπαίγματος θα έχουν συνεπέστερη παρουσία, βελτιώνοντας την ικανότητα του πράκτορα να αναλύει και να ανταποκρίνεται στο περιβάλλον.



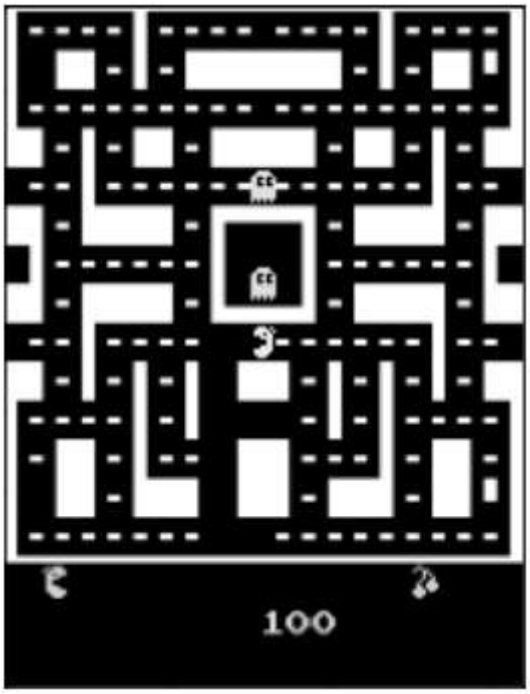
Εικόνα 6. Δύο διαδοχικές καταστάσεις του περιβάλλοντος του παιχνιδιού, όπου τα εικονοστοιχεία του κόκκινου και του μοβ φαντάσματος τρεμοπαίζουν.

Διαμόρφωση καρέ

Η εισαγωγή εικόνας που λαμβάνει ο πράκτορας αποτελείται από τα δεδομένα που προέρχονται από την μηχανή Arcade, και η πολυπλοκότητα της εκπαίδευσης του πράκτορα συνδέεται άμεσα με τις πληροφορίες που παρέχονται μέσω αυτών των εισόδων. Οι εισαγωγές απεικονίζονται ως τιμές ενός τρισδιάστατου πίνακα.

Για να μειωθεί αυτή η διαστασιμότητα και να βελτιωθεί η επεξεργαστική απόδοση, εφαρμόζονται

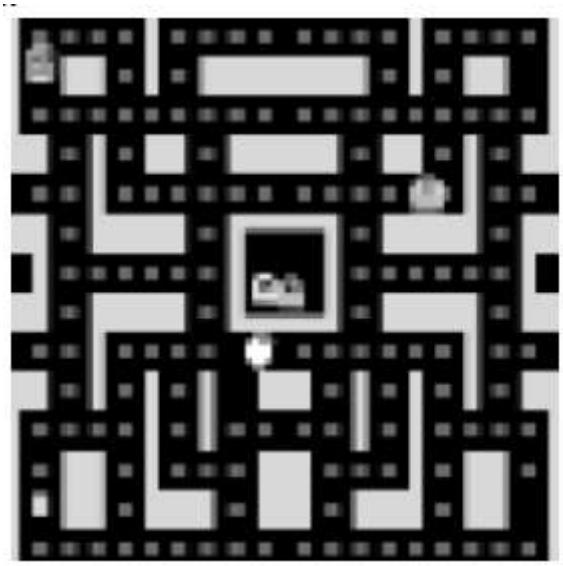
μετασχηματισμοί στα δεδομένα εισόδου που δέχεται ο πράκτορας. Αρχικά, το οπτικό σήμα μετατρέπεται από το πολύχρωμο φάσμα RGB σε μονόχρωμη κλίμακα του γκρι (Gray-scale), μειώνοντας τις διαστάσεις των δεδομένων. Αυτός ο μετασχηματισμός, που υλοποιείται με τη βοήθεια της βιβλιοθήκης OpenCV, επιτρέπει στον πράκτορα να επεξεργάζεται τις πληροφορίες με μεγαλύτερη ταχύτητα και αποτελεσματικότητα.



Εικόνα 7. Ένα καρέ του παιχνιδιού σε Grayscale.

Επιπρόσθετα, οι διαστάσεις της εικόνας αλλάζουν προκειμένου να βελτιωθεί η απόδοση των συνελικτικών νευρωνικών δικτύων, τα οποία αποδίδουν καλύτερα με τετράγωνες εικόνες. Με την αναμόρφωση της εικόνας σε διαστάσεις (110,84,1) με τη χρήση της μεθόδου INTER_AREA, πραγματοποιείται μια πιο αποδοτική εκμετάλλευση της σχέσης ανάμεσα στα εικονοστοιχεία.

Τέλος, τα μέρη της εικόνας που περιέχουν αχρείαστες πληροφορίες, όπως τα σκορ και ο αριθμός των ζωών, περικόπτονται για να για να μπορέσει να εκπαιδευτεί ο πράκτορας στις πραγματικά σημαντικές πληροφορίες του περιβάλλοντος. Αυτή η στοχευμένη προσέγγιση εξασφαλίζει ότι οι διαθέσιμοι υπολογιστικοί πόροι χρησιμοποιούνται με τον πιο αποτελεσματικό τρόπο, ενισχύοντας την ολοκληρωμένη απόδοση του μοντέλου μάθησης.



Εικόνα 8. Περικομμένο καρέ.

Για την δημιουργία του παιχνιδιού θα γίνει χρήση της βιβλιοθήκης gym της openai.

Κατασκευή πλειάδων καρέ (Frame stacking)

Κατασκευάζοντας πλειάδες καρέ (frame stacking), βελτιώνεται η αντίληψη του πράκτορα για τη δυναμική του περιβάλλοντος σε ένα παιχνίδι. Κάθε μεμονωμένο καρέ παρέχει περιορισμένη ενημέρωση, επιδεικνύοντας μόνο τις θέσεις των φιγούρων και των αντιπάλων, χωρίς να καταγράφει τις κινήσεις τους. Η δημιουργία ενός συνόλου από διαδοχικά καρέ, συνήθως τέσσερα, παρέχει μια πληρέστερη εικόνα της δυναμικής του παιχνιδιού, καθώς ο πράκτορας μπορεί να παρατηρήσει την κίνηση στον χώρο.

Η διαδικασία αυτή επιτρέπει στον πράκτορα να παρακολουθεί τις κινήσεις των φαντασμάτων και άλλων δυναμικών στοιχείων του παιχνιδιού. Αυτές οι διαδοχικές παρατηρήσεις διαφέρουν ελάχιστα μεταξύ τους, αφού κάθε νέα πλειάδα αντικαθιστά μόνο ένα καρέ από το προηγούμενο σύνολο. Λόγω του ότι απαιτείται η αποθήκευση ενός μεγάλου όγκου τέτοιων πλειάδων, χρησιμοποιείται η τεχνική LazyFrames, η οποία εξοικονομεί μνήμη αποθηκεύοντας μόνο μία φορά τα καρέ για κάθε διαδοχικό σύνολο.

4.3 **Πράκτορας**

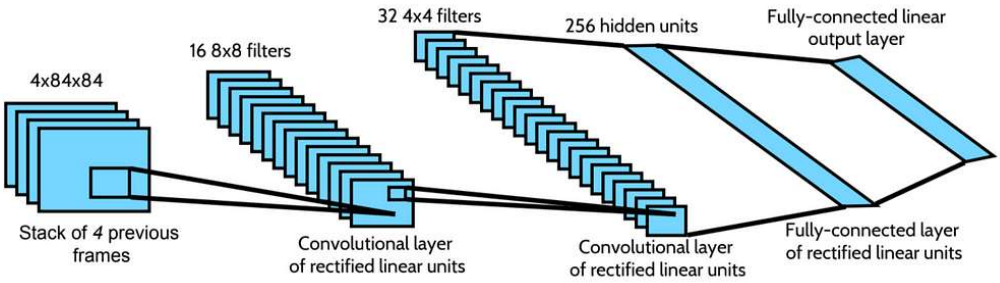
Στο επίκεντρο της συγκεκριμένης εφαρμογής είναι η δημιουργία και η ανάπτυξη ενός πράκτορα που βασίζεται στη βαθιά ενισχυτική μάθηση. Αυτός ο πράκτορας εκπαιδεύεται να αλληλεπιδρά με το περιβάλλον του παιχνιδιού MsPacman, στοχεύοντας στη συλλογή των μεγαλύτερων δυνατών ανταμοιβών σε κάθε επεισόδιο. Αυτή η διαδικασία υλοποιείται με ένα νευρωνικό δίκτυο, το οποίο λει-

τουργεί ως ο πράκτορας, με τα βάρη του να καθορίζουν και να διαμορφώνουν τη συμπεριφορά του εντός του παιχνιδιού.

Πράκτορας DQN

Το νευρωνικό δίκτυο DQN (Deep Q Network) περιλαμβάνει ένα συνελικτικό δίκτυο με τρία διαδοχικά συνελικτικά επίπεδα και δύο πλήρως συνδεδεμένα επίπεδα που τα ακολουθούν. Αρχικά, το πρώτο συνελικτικό επίπεδο δέχεται τις καταστάσεις του περιβάλλοντος όπως έχουν προεπεξεργαστεί, χρησιμοποιώντας 32 φίλτρα με διαστάσεις 8x8 και βήμα 4. Το δεύτερο συνελικτικό επίπεδο αποτελείται από 64 φίλτρα 4x4 με βήμα 2, ενώ το τρίτο και τελευταίο συνελικτικό επίπεδο χρησιμοποιεί επίσης 64 φίλτρα, αλλά με διαστάσεις 3x3 και μοναδιαίο βήμα σάρωσης.

Το επόμενο επίπεδο, που είναι πλήρως συνδεδεμένο, περιλαμβάνει 512 νευρώνες και ακολουθείται από ένα τελικό πλήρως συνδεδεμένο επίπεδο, το οποίο έχει αριθμό νευρώνων ίσο με τις επιτρεπόμενες κινήσεις του πράκτορα στο παιχνίδι, οι οποίες στο MsPacman είναι εννέα. Κάθε επίπεδο, εκτός από το τελευταίο, ενσωματώνει τη μη-γραμμική συνάρτηση ενεργοποίησης ReLu για την προσθήκη μη-γραμμικότητας στην επεξεργασία της πληροφορίας. Η έξοδος του δικτύου αντιπροσωπεύει τις τιμές του πίνακα Q(s,a), παρέχοντας τις αξίες για κάθε δυνατή δράση ανάλογα με την εισαγόμενη κατάσταση.



Εικόνα 9: Η αρχιτεκτονική του δικτύου DQN

Πηγή: https://jonathan-hui.medium.com/rl-dqn-deep-q-network-e207751f7ae4

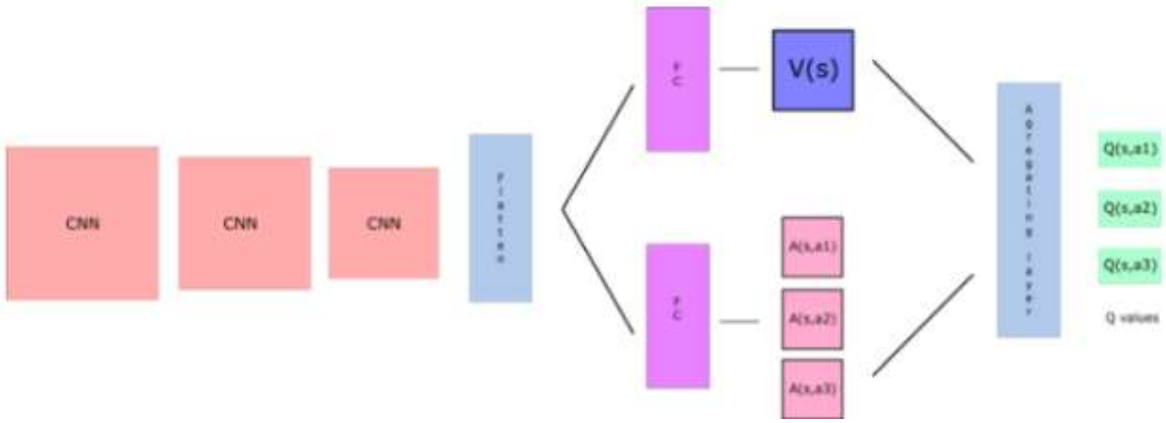
Παρόμοια αρχιτεκτονική χρησιμοποιήθηκε και για τους πράκτορες double DQN και noisy DQN. Στον noisy DQN, τα τελευταία πλήρως συνδεδεμένα επίπεδα, έχουν αντικατασταθεί από πλήρως συνδεδεμένα θορυβώδη επίπεδα.

Πράκτορας Duel DQN

Το Duel DQN (Duel Q Network) είναι μια παραλλαγή του Deep Q Network, που σχεδιάστηκε για να βελτιώσει την εκτίμηση των πράξεων που πρέπει να πραγματοποιήσει ο πράκτορας. Η βασική αρχιτεκτονική του Duel DQN απαιτεί τροποποίηση του κλασικού DQN με την προσθήκη δύο παράλληλων πλήρως συνδεδεμένων επιπέδων μετά το τελευταίο συνελικτικό επίπεδο.

Το πρώτο από αυτά τα επίπεδα περιλαμβάνει 512 νευρώνες και χρησιμοποιεί τη μη-γραμμική συνάρτηση ενεργοποίησης ReLu για να επιτρέπει την επεξεργασία πληροφορίας με μη-γραμμικό τρόπο. Το δεύτερο επίπεδο περιέχει μόλις ένα νευρώνα και δεν χρησιμοποιεί συνάρτηση ενεργοποίησης, εξυπηρετώντας στην άμεση εκτίμηση της τιμής της κατάστασης, ανεξάρτητα από τις διαθέσιμες πράξεις στον πράκτορα.

Αυτή η δομή επιτρέπει στο Duel DQN να αξιολογεί ξεχωριστά την αξία της καλύτερης πράξης που μπορεί να γίνει σε μια συγκεκριμένη κατάσταση (διαμέσου του ενός νευρώνα) και το πλεονέκτημα κάθε πράξης σε σχέση με τις άλλες (διαμέσου των 512 νευρώνων), ενισχύοντας την απόδοση του πράκτορα σε σύνθετα περιβάλλοντα.



Εικόνα 10. Το δίκτυο Duel DQN.

Πηγή: Improvements in Deep Q Learning: Dueling Double DQN, Prioritized Experience Replay, and fixed… (freecodecamp.org)

Experience Replay Buffer

Ο αλγόριθμος που χρησιμοποιείται για την εκπαίδευση του πράκτορα ενισχυτικής μάθησης βασίζεται στη δημιουργία και χρήση ενός "Experience Replay Buffer". Αυτός ο buffer αποτελεί έναν τύπο μνήμης, στην οποία αποθηκεύονται οι μεταβάσεις του πράκτορα στη μορφή (s, a, r, s', done), όπου s είναι η κατάσταση, a η δράση, r η ανταμοιβή, s' η νέα κατάσταση, και done δηλώνει εάν η επεισοδιακή διαδικασία έχει τελειώσει.

Η μνήμη αυτή λειτουργεί με την αρχή FIFO (First In First Out) όταν το συνολικό πλήθος των μεταβάσεων υπερβεί τον καθορισμένο χώρο αποθήκευσης. Σε περιόδους εκπαίδευσης, επιλέγονται τυχαία δείγματα από τον buffer με σκοπό την ομοιόμορφη κατανομή, ορίζοντας το batch_size ως το πλήθος των μεταβάσεων που θα χρησιμοποιηθούν σε κάθε βήμα της εκπαίδευσης. Κάθε μετάβαση καταλαμβάνει έναν χώρο στη μνήμη και οι πληροφορίες από όλες τις επιλεγμένες μεταβάσεις συγκεντρώνονται σε κοινούς πίνακες, που αποτελούν την έξοδο του buffer.

Η δομή και η διαχείριση αυτής της μνήμης επιτυγχάνεται μέσω της κλάσης Replay Buffer, η οποία διαθέτει τις απαραίτητες συναρτήσεις για την αποθήκευση και την δειγματοληψία των μεταβάσεων, διασφαλίζοντας την ομαλή λειτουργία και την αποτελεσματική χρήση της μνήμης για την εκπαίδευση του πράκτορα.

Εκπαίδευση πράκτορα

Η διαδικασία εκπαίδευσης του πράκτορα ενισχυτικής μάθησης περιλαμβάνει τακτική ανανέωση των βαρών του νευρωνικού δικτύου κάθε ορισμένο αριθμό βημάτων, ο οποίος καθορίζεται από την παράμετρο update_frequency. Σε κάθε βήμα εκπαίδευσης, ένα υποσύνολο μεταβάσεων από τη μνήμη

(Experience Replay Buffer) μεταφέρεται στην κάρτα γραφικών για επεξεργασία. Αυτές οι μεταβάσεις επεξεργάζονται για να παράγουν τις τιμές εισόδου για τη συνάρτηση σφάλματος, η οποία στην περίπτωση αυτή είναι η Huber Loss (ή smooth_l1_loss). Η Huber Loss είναι προτιμητέα σε σημεία όπου οι εκτιμήσεις Q είναι πολύ θορυβώδεις και παρέχει έναν ομαλότερο υπολογισμό του σφάλματος.

Από τη συνάρτηση σφάλματος προκύπτει το σφάλμα που χρησιμοποιείται για την οπισθοδρομική διάδοση (backpropagation), όπου υπολογίζονται οι μερικές παράγωγοι του σφάλματος ως προς κάθε βάρος του δικτύου. Για να προστεθεί επιπλέον σταθερότητα στη διαδικασία εκπαίδευσης, οι μερικές παραγώγους ψαλιδίζονται στο διάστημα (-1,1).

Για την εκτέλεση της βελτιστοποίησης των βαρών, χρησιμοποιείται ο αλγόριθμος Adam. Ο αλγόριθμος Adam δέχεται τις ψαλιδισμένες μερικές παραγώγους του σφάλματος και προχωρά στην ανανέωση των βαρών βάσει ενός ορισμένου ρυθμού μάθησης (learning rate, LR). Τέλος, για την περαιτέρω βελτίωση της σταθερότητας και της απόδοσης του δικτύου, τα βάρη του στόχου δικτύου ανανεώνονται με χρήση του αλγορίθμου soft update με ένα ρυθμό TAU, διασφαλίζοντας μια ομαλή μετάβαση των μαθημένων γνώσεων στο δίκτυο.

5 Αποτελέσματα

Η εκπαίδευση του πράκτορα πραγματοποιήθηκε στο Google Colab. Για την επιτάχυνση των υπολογισμών χρησιμοποιήθηκε η T4 GPU με 12.7GB συστημική RAM και 15GB GPU RAM. Επειδή η RAM ήταν ανεπαρκής σε μεγάλο αριθμό βημάτων, η εκπαίδευση έγινε για 100.000 βήματα, που αντιστοιχεί σε περίπου 50 επεισόδια. Ως μετρικές για την απόδοση του κάθε αλγόριθμου χρησιμοποιήθηκε η μέση ανταμοιβή (reward) των τελευταίων δύο επεισοδίων, η ανταμοιβή κάθε επεισοδίου (score) και η ταχύτητα του αλγόριθμου που υπολογίζεται από τον αριθμό των βημάτων ανά δευτερόλεπτο.

Οι τιμές των μεταβλητών της εκπαίδευσης δίνονται στον παρακάτω πίνακα

Μεταβλητή	**DQN**	**Double DQN**	**Duel DQN**	**Noisy DQN**
episodic_life	True	True	True	True
Clip_rewards	True	True	True	True
Use_double_dqn	False	True	False	False
Use_dueling	False	False	True	False
Use_noisy	False	False	False	True
Use_normalized	False	True	False	False

Πίνακας 2 Οι τιμές των μεταβλητών της εκπαίδευσης.

Οι υπερπαράμετροι του πειράματος παρουσιάζονται στον παρακάτω πίνακα.

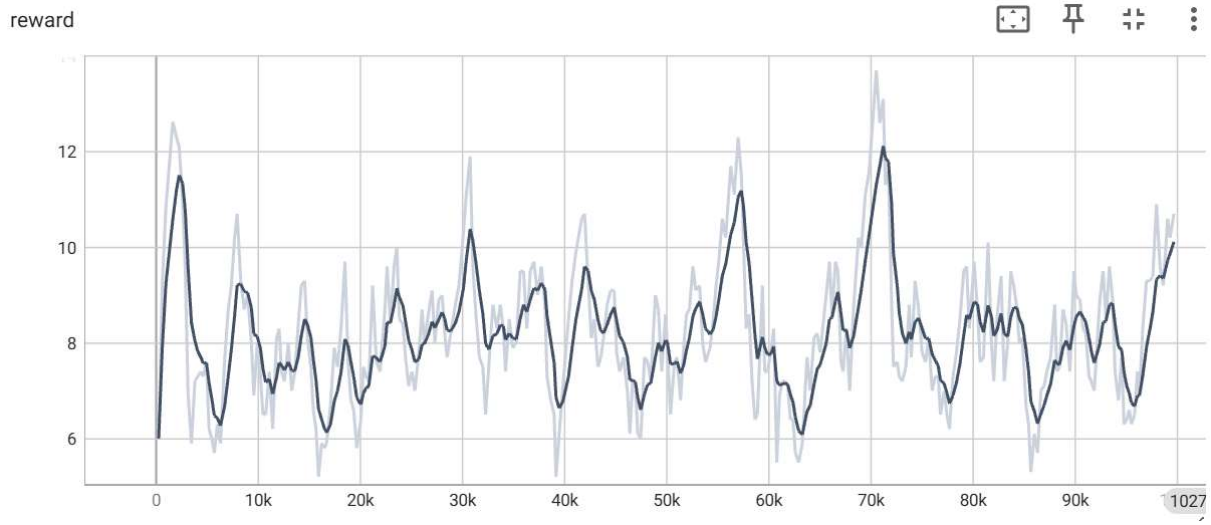
Υπερπαράμετρος	Τιμή
batch size	32
memory size	50000
frame stacking	4
frame skipping	4
discount factor (y)	0.99
target update	10000
update frequency	4
learning rate(lir)	0.00025
eps_start	1
eps end	0.01
eps decay	1000000
max frames	10000
min frames	500
no-op max	30

Πίνακας 3 Οι υπερπαράμετροι του πειράματος

5.1 Πράκτορας DQN

Στην Εικόνα 11 έχουμε την μέση ανταμοιβή του πράκτορα double DQN. Η ανταμοιβή παρουσιάζει σημαντικές διακυμάνσεις μεταξύ των επεισοδίων, κάτι που υποδηλώνει μεταβλητότητα στην απόδοση του πράκτορα. Στην αρχική περίοδο της εκπαίδευσης (0-10k επεισόδια) παρατηρείται μια απότομη αύξηση στη μέση ανταμοιβή, φτάνοντας σε μέγιστη τιμή περίπου 12 πριν αρχίσει η πτώση. Αυτή η αύξηση είναι συνήθης καθώς ο πράκτορας μαθαίνει βασικές στρατηγικές για το παιχνίδι.

Στην συνέχεια παρατηρείται σημαντική διακύμανση στη μέση ανταμοιβή. Αυτές οι αυξομειώσεις υποδεικνύουν τις δοκιμές και τις προσαρμογές του πράκτορα καθώς ανακαλύπτει και αξιολογεί διάφορες στρατηγικές. Οι αιχμές και οι πτώσεις είναι αποτέλεσμα της διαδικασίας διερεύνησης και εκμετάλλευσης.

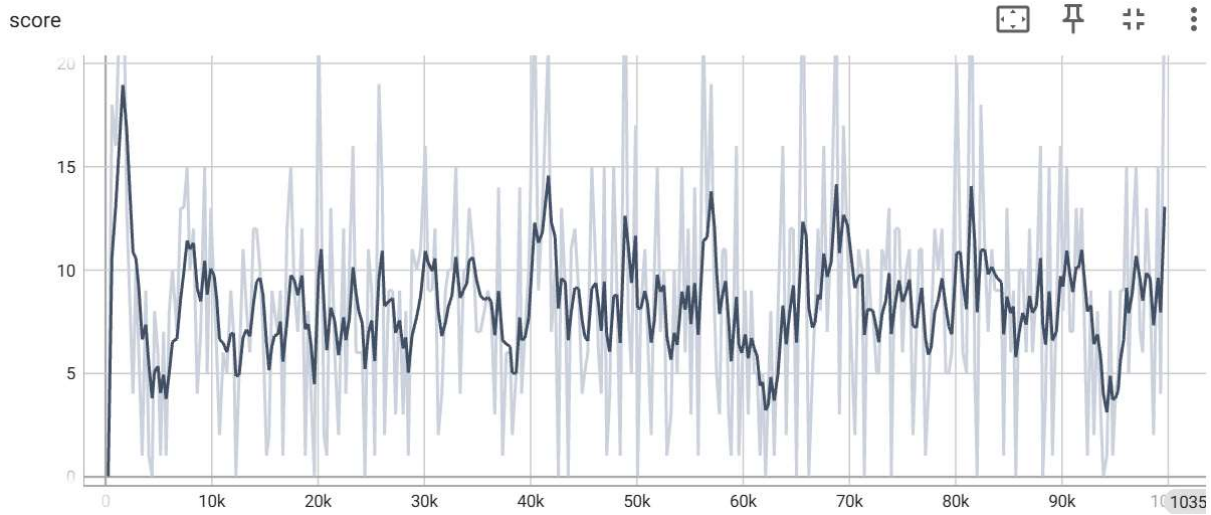


Εικόνα 11 Η μέση ανταμοιβή για κάθε δέκα επεισόδια για τον double DQN.

Στην Εικόνα 12 έχουμε το σκορ του πράκτορα double DQN σε κάθε γύρο. Υπάρχουν αρκετές κορυφές και χαμηλά σημεία, γεγονός που δείχνει ότι η απόδοση του πράκτορα δεν είναι σταθερή.

Στην αρχική περίοδο της εκπαίδευσης παρατηρείται μια απότομη αύξηση στο σκορ, φτάνοντας σε μέγιστη τιμή πάνω από 25 πριν αρχίσει η πτώση. Αυτή η φάση μπορεί να αποδοθεί στην διαδικασία εξερεύνησης, όπου ο πράκτορας προσπαθεί να μάθει την στρατηγική του παιχνιδιού μέσω τυχαίων ενεργειών και ανατροφοδότησης.

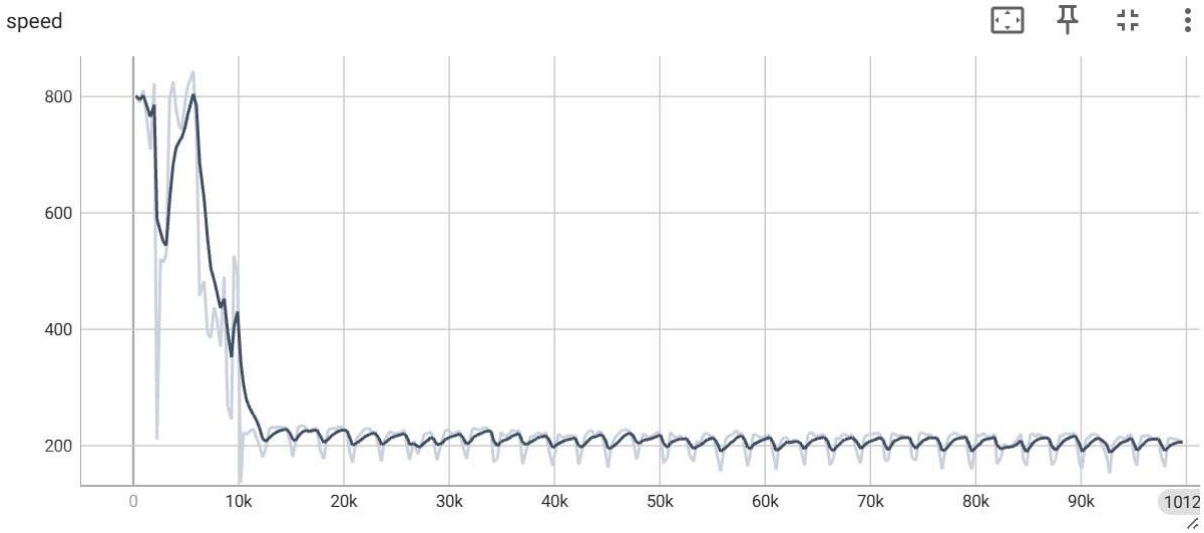
Στην συνέχεια παρατηρείται σημαντική διακύμανση στο σκορ. Αυτές οι αυξομειώσεις υποδεικνύουν τις δοκιμές και τις προσαρμογές του πράκτορα καθώς ανακαλύπτει και αξιολογεί διάφορες στρατηγικές. Οι αιχμές και οι πτώσεις είναι αποτέλεσμα της διαδικασίας διερεύνησης και εκμετάλλευσης.



Εικόνα 12 Το σκορ του πράκτορα DQN.

Στην Εικόνα 13 έχουμε την ταχύτητα εκπαίδευσης του πράκτορα DQN. Η ταχύτητα του πράκτορα κατά τη διάρκεια της εκπαίδευσης δεν είναι σταθερή. Στην αρχική περίοδο της εκπαίδευσης παρατηρείται υψηλή ταχύτητα εκπαίδευσης, φτάνοντας μέχρι και 800 βήματα/δευτερόλεπτο.
Η ταχύτητα μειώνεται γρήγορα στα πρώτα 10k επεισόδια καθώς το σύστημα αρχίζει να προσαρμόζεται στις απαιτήσεις του περιβάλλοντος και της εκπαίδευσης.
Μετά την αρχική περίοδο, η ταχύτητα σταθεροποιείται γύρω στα 200 βήματα/δευτερόλεπτο.
Αυτή η σταθεροποίηση μπορεί να υποδηλώνει ότι ο πράκτορας έχει φτάσει σε μια πιο σταθερή κατάσταση λειτουργίας.

Η ταχύτητα εκπαίδευσης διατηρείται σταθερή γύρω στα 200 βήματα/δευτερόλεπτο για το υπόλοιπο της εκπαίδευσης. Η σταθεροποίηση υποδηλώνει ότι το σύστημα έχει βρει έναν πιο αποδοτικό ρυθμό επεξεργασίας των πληροφοριών καθώς προχωρά η εκπαίδευση. Οι μικρές διακυμάνσεις που παρατηρούνται κατά τη διάρκεια της εκπαίδευσης είναι αναμενόμενες και αντανακλούν την προσαρμοστική φύση της διαδικασίας ενισχυτικής μάθησης.



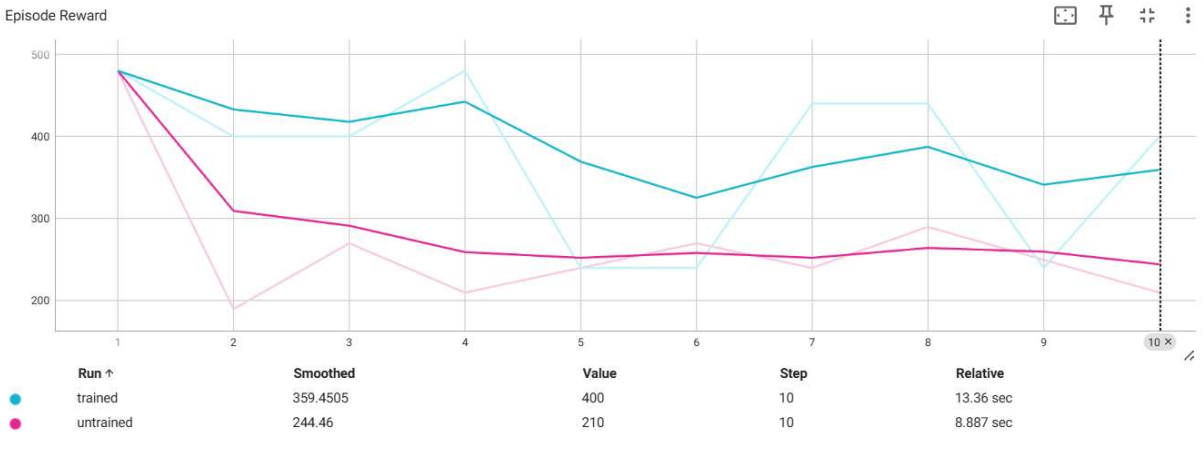
Εικόνα 13 Η ταχύτητα εκπαίδευσής του πράκτορα DQN.

Στην Εικόνα 14 παρουσιάζεται η σύγκριση των ανταμοιβών που επιτυγχάνονται από ένα εκπαιδευμένο μοντέλο και ένα ανεκπαίδευτο μοντέλο DQN κατά τα πρώτα 10 επεισόδια. Το διάγραμμα αποτυπώνει την επίδοση των δύο μοντέλων σε όρους ανταμοιβής ανά επεισόδιο.

Το εκπαιδευμένο μοντέλο ξεκινά με υψηλότερη ανταμοιβή (περίπου 500) σε σύγκριση με το ανεκπαίδευτο μοντέλο (περίπου 400). Αυτή η διαφορά υποδεικνύει ότι το εκπαιδευμένο μοντέλο έχει ήδη μάθει αποτελεσματικές στρατηγικές από την εκπαίδευση.

Στην συνέχεια, το εκπαιδευμένο μοντέλο δείχνει σταδιακή σταθεροποίηση και βελτίωση των επιδόσεων, με τις ανταμοιβές να κυμαίνονται γύρω στα 350-400. Το ανεκπαίδευτο μοντέλο παραμένει χαμηλότερο στις ανταμοιβές, με μικρές αυξομειώσεις, υποδεικνύοντας αδυναμία να αναπτύξει αποτελεσματικές στρατηγικές μέσα σε αυτά τα επεισόδια.

Η σύγκριση των ανταμοιβών για τα πρώτα 10 επεισόδια ανάμεσα στο εκπαιδευμένο και το ανεκπαίδευτο μοντέλο DQN δείχνει ξεκάθαρα την υπεροχή του εκπαιδευμένου μοντέλου. Το εκπαιδευμένο μοντέλο ξεκινά με υψηλότερη ανταμοιβή και διατηρεί σταθερά καλύτερες επιδόσεις σε σχέση με το ανεκπαίδευτο μοντέλο καθ' όλη τη διάρκεια των πρώτων 10 επεισοδίων.

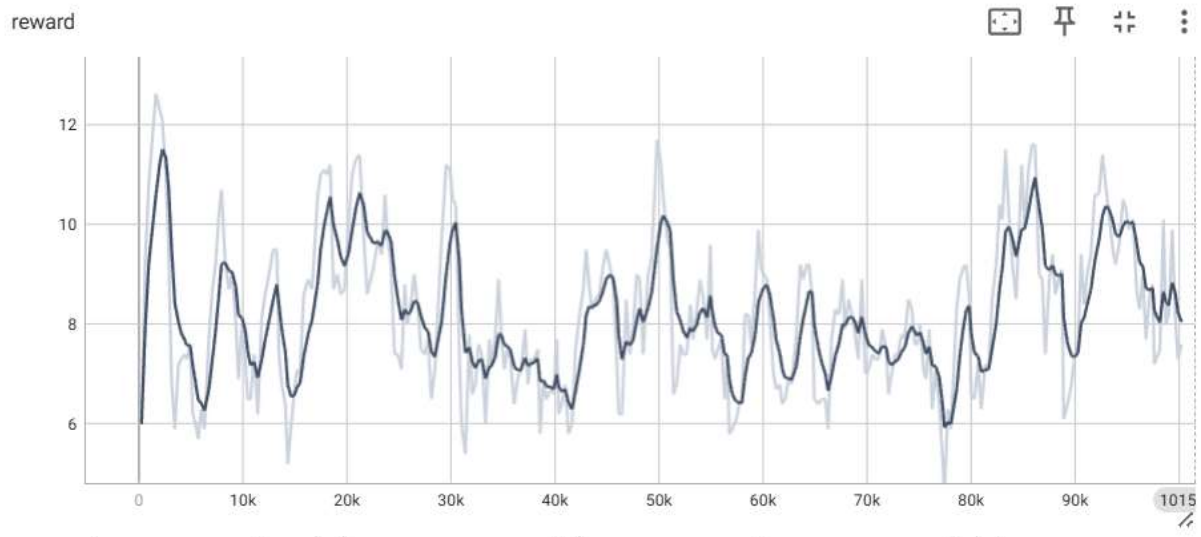


Εικόνα 14 Σύγκριση του εκπαιδευμένου πράκτορα DQN με έναν ανεκπαίδευτο.

5.2 Πράκτορας DQN με batch normalization

Στην Εικόνα 15 έχουμε την μέση ανταμοιβή για κάθε δέκα επεισόδια του πράκτορα double DQN με batch normalization. Η αρχική περίοδος δείχνει απότομη αύξηση στη μέση ανταμοιβή, φτάνοντας σε μέγιστη τιμή πάνω από 12 πριν αρχίσει η πτώση.
Αυτή η γρήγορη άνοδος είναι χαρακτηριστική καθώς ο πράκτορας με Batch Normalization προσαρμόζεται γρήγορα στο περιβάλλον.

Στην συνέχεια παρατηρείται σημαντική διακύμανση στη μέση ανταμοιβή με αυξομειώσεις. Η χρήση του Batch Normalization φαίνεται να βοηθά στη σταθεροποίηση της απόδοσης του πράκτορα, αν και υπάρχουν πτώσεις σε ορισμένες περιπτώσεις. Η μέση ανταμοιβή κυμαίνεται συνήθως μεταξύ 6 και 10, με περιοδικές κορυφές και πτώσεις.

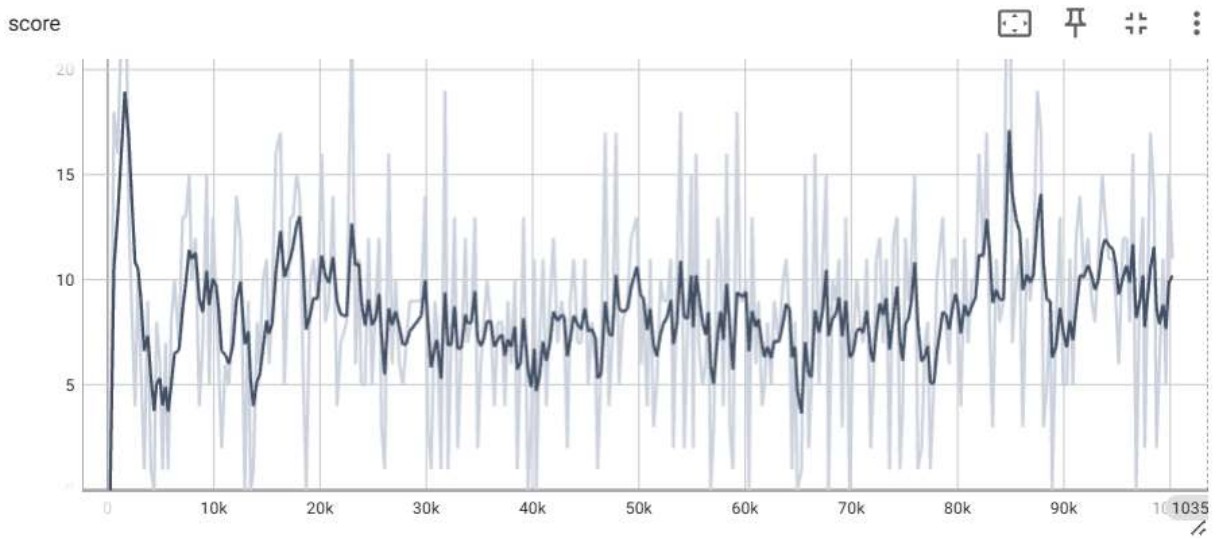


Εικόνα 15 Η μέση ανταμοιβή για κάθε δέκα επεισόδια για τον πράκτορα double DQN με batch normalization.

Στην Εικόνα 16 παρουσιάζεται το σκορ που επιτυγχάνεται από τον πράκτορα DQN με χρήση Batch Normalization κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Το διάγραμμα αποτυπώνει τη διακύμανση του σκορ με την πρόοδο της εκπαίδευσης, υπολογίζοντας το σκορ για κάθε επεισόδιο.

Στην αρχική περίοδο της εκπαίδευσης παρατηρείται μια απότομη αύξηση στο σκορ, φτάνοντας μέχρι και 20 πριν αρχίσει η πτώση. Στην συνέχεια υπάρχει σημαντική διακύμανση στο σκορ. Οι διακυμάνσεις είναι αποτέλεσμα της διαδικασίας διερεύνησης και εκμετάλλευσης που ακολουθεί ο πράκτορας κατά τη διάρκεια της εκπαίδευσης.

Η χρήση Batch Normalization στο DQN φαίνεται να μειώνει τη μεταβλητότητα, προσφέροντας μια πιο σταθερή απόδοση.

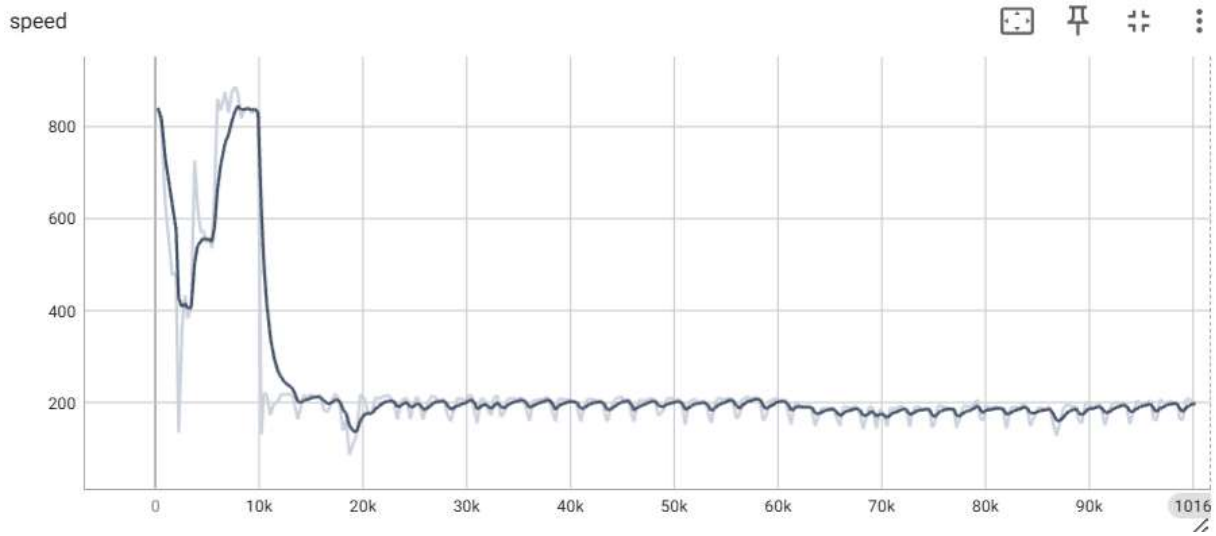


Εικόνα 16 Το σκορ του πράκτορα DQN με batch normalization.

Στην εικόνα 17 παρουσιάζεται η ταχύτητα εκπαίδευσης του πράκτορα DQN με χρήση Batch Normalization σε όρους βήματα/δευτερόλεπτο κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Το διάγραμμα αποτυπώνει την απόδοση του συστήματος με την πάροδο του χρόνου, δίνοντας μια εικόνα για την αποδοτικότητα της διαδικασίας εκπαίδευσης.

Στην αρχική περίοδο της εκπαίδευσης παρατηρείται υψηλή ταχύτητα εκπαίδευσης, φτάνοντας μέχρι και πάνω από 800 βήματα/δευτερόλεπτο.
Η ταχύτητα μειώνεται στα πρώτα 10k επεισόδια καθώς το σύστημα αρχίζει να προσαρμόζεται στις απαιτήσεις του περιβάλλοντος και της εκπαίδευσης.

Η ταχύτητα εκπαίδευσης του πράκτορα DQN με Batch Normalization παρουσιάζει σημαντική μείωση κατά τα αρχικά στάδια της εκπαίδευσης, η οποία σταθεροποιείται στη συνέχεια. Η αρχική υψηλή ταχύτητα είναι χαρακτηριστική της έναρξης της εκπαίδευσης, όπου ο πράκτορας επεξεργάζεται γρήγορα τις αρχικές πληροφορίες και προσαρμόζεται στις βασικές στρατηγικές του παιχνιδιού. Η σταθεροποίηση γύρω στα 200 βήματα/δευτερόλεπτο υποδηλώνει ότι το σύστημα έχει βρει έναν πιο αποδοτικό ρυθμό επεξεργασίας καθώς προχωρά η εκπαίδευση. Οι μικρές διακυμάνσεις που παρατηρούνται κατά τη διάρκεια της εκπαίδευσης είναι αναμενόμενες και αντανακλούν την προσαρμοστική φύση της διαδικασίας ενισχυτικής μάθησης.

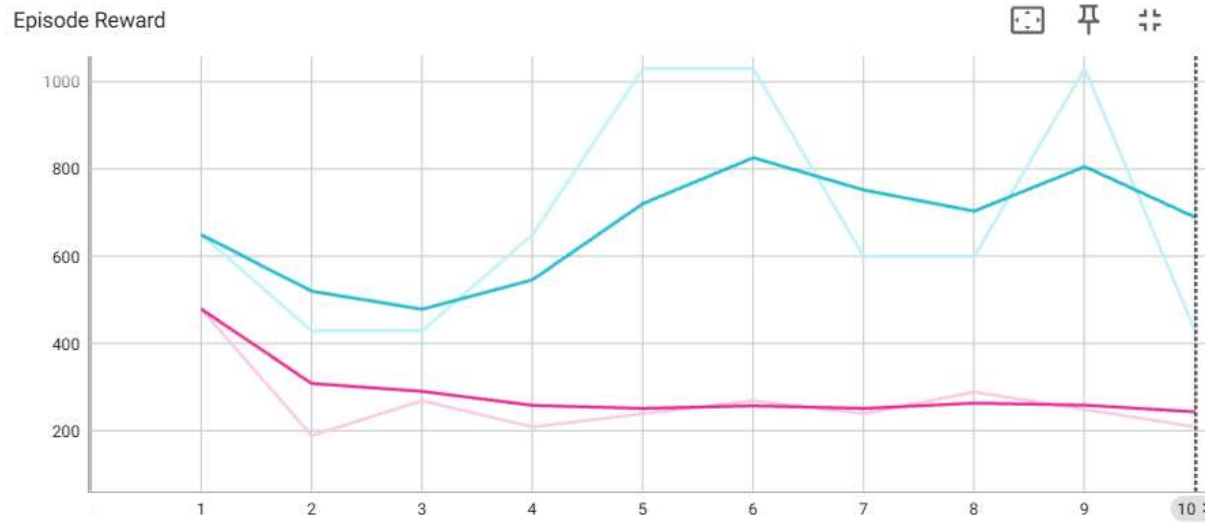


Εικόνα 17 Η ταχύτητα εκπαίδευσής του πράκτορα DQN με batch normalization.

Στην εικόνα 18 παρουσιάζεται η σύγκριση των ανταμοιβών που επιτυγχάνονται από ένα εκπαιδευμένο και ένα ανεκπαίδευτο μοντέλο DQN κατά τα πρώτα 10 επεισόδια. Το διάγραμμα αποτυπώνει την επίδοση των δύο μοντέλων σε όρους ανταμοιβής ανά επεισόδιο.

Το εκπαιδευμένο μοντέλο ξεκινά με υψηλότερη ανταμοιβή (περίπου 800) σε σύγκριση με το ανεκπαίδευτο μοντέλο (περίπου 400). Αυτή η διαφορά υποδεικνύει ότι το εκπαιδευμένο μοντέλο έχει ήδη μάθει αποτελεσματικές στρατηγικές από την εκπαίδευση.

Η σύγκριση των ανταμοιβών για τα πρώτα 10 επεισόδια ανάμεσα στο εκπαιδευμένο και το ανεκπαίδευτο μοντέλο DQN δείχνει ξεκάθαρα την υπεροχή του εκπαιδευμένου μοντέλου. Το εκπαιδευμένο μοντέλο ξεκινά με υψηλότερη ανταμοιβή και διατηρεί σταθερά καλύτερες επιδόσεις σε σχέση με το ανεκπαίδευτο μοντέλο καθ' όλη τη διάρκεια των πρώτων 10 επεισοδίων.



Run ↑	Smoothed	Value	Step	Relative
trained	689.7182	430	10	15.03 sec
untrained	244.46	210	10	7.954 sec

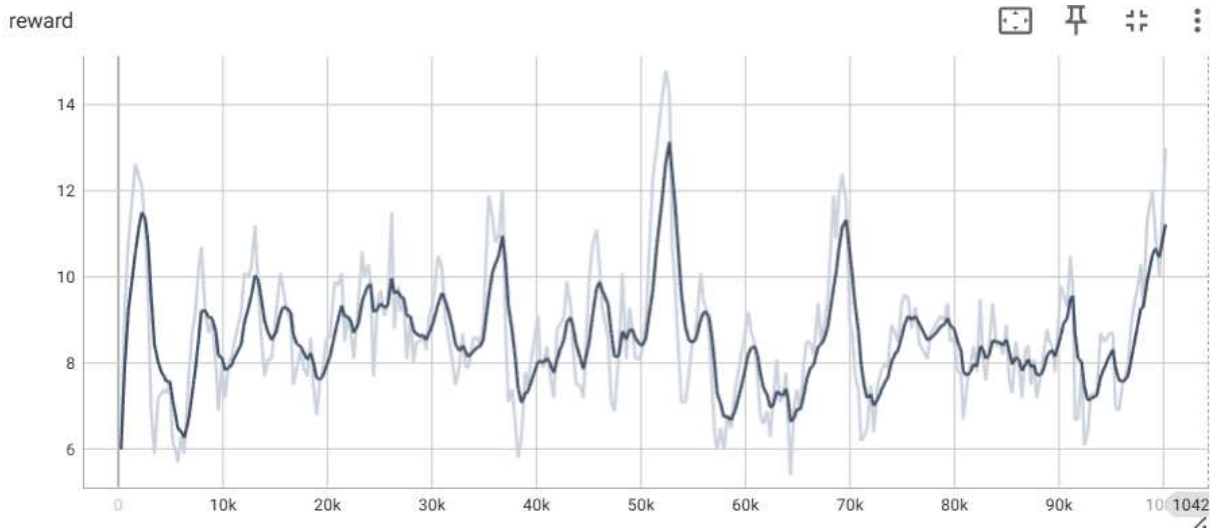
Εικόνα 18Σύγκριση του εκπαιδευμένου πράκτορα DQN με batch normalization με έναν ανεκπαίδευτο.

5.3 Πράκτορας duel DQN

Στην εικόνα 19 παρουσιάζεται η μέση ανταμοιβή που επιτυγχάνεται από τον πράκτορα Dueling DQN κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Το διάγραμμα αποτυπώνει τη διακύμανση της ανταμοιβής με την πρόοδο της εκπαίδευσης, υπολογίζοντας τη μέση ανταμοιβή για κάθε δέκα επεισόδια

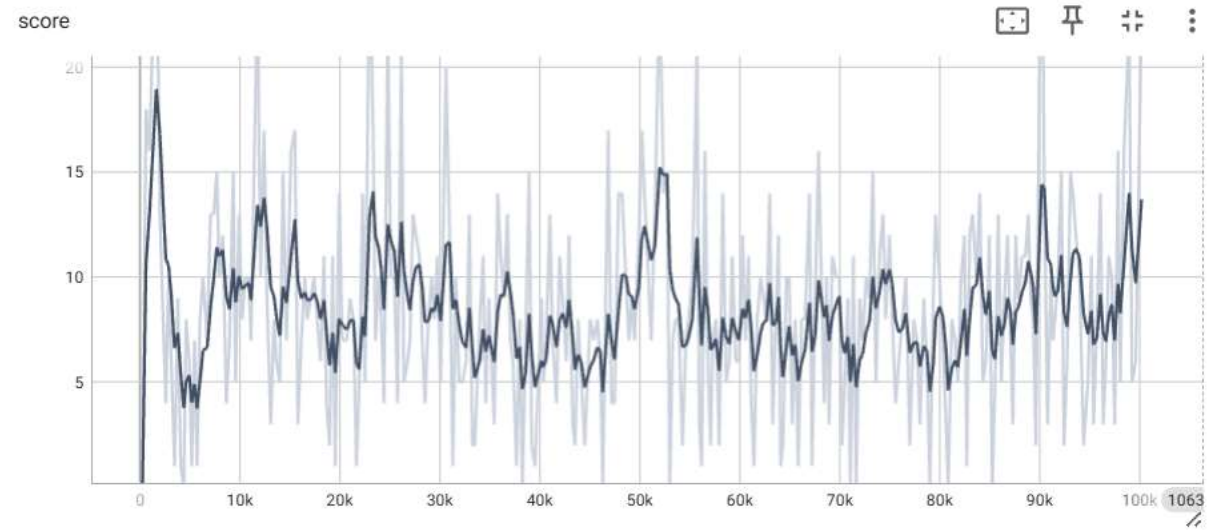
Στην αρχική περίοδο της εκπαίδευσης παρατηρείται απότομη αύξηση στη μέση ανταμοιβή, φτάνοντας μέχρι και πάνω από 12 πριν αρχίσει η πτώση. Στην συνέχεια η μέση ανταμοιβή δείχνει μια τάση προς σταθεροποίηση, αν και με συνεχιζόμενη μεταβλητότητα. Παρατηρούνται αιχμές προς το τέλος της εκπαίδευσης που φτάνουν έως και 11.

Το Dueling DQN διατηρεί την σταθερότητα της απόδοσης και επιτυγχάνει υψηλότερες μέσες ανταμοιβές, υποδεικνύοντας βελτιωμένη μάθηση και εκμετάλλευση των στρατηγικών.



Εικόνα 19 Η μέση ανταμοιβή για κάθε δύο επεισόδια για τον πράκτορα Duel DQN.

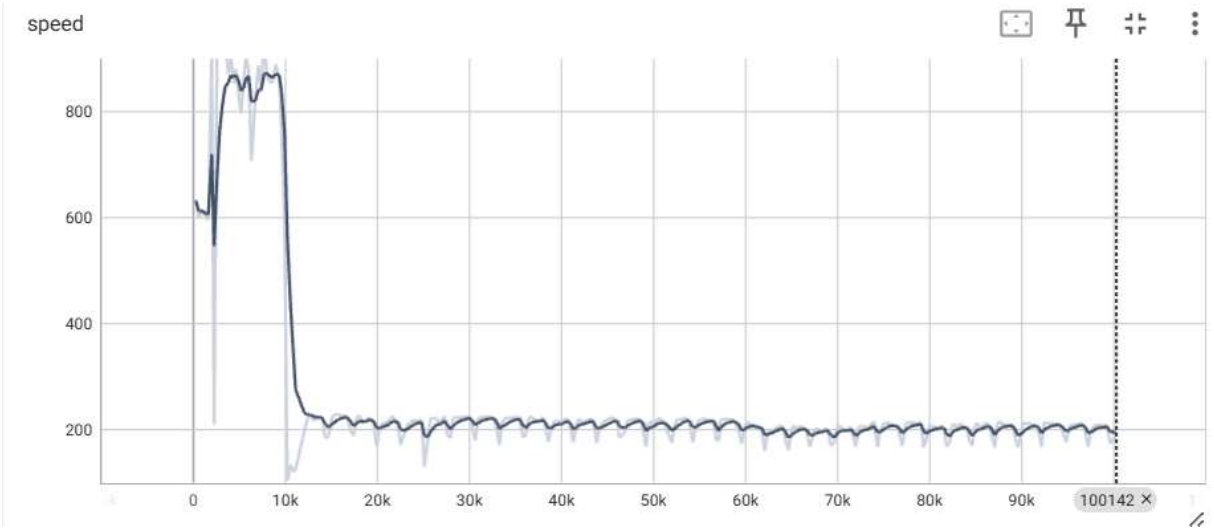
Στην Εικόνα 20 παρουσιάζεται το σκορ που επιτυγχάνεται από τον πράκτορα Dueling DQN κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Το διάγραμμα αποτυπώνει τη διακύμανση του σκορ με την πρόοδο της εκπαίδευσης, υπολογίζοντας το σκορ για κάθε επεισόδιο. Στην αρχική περίοδο της εκπαίδευσης παρατηρείται απότομη αύξηση στο σκορ, φτάνοντας μέχρι και πάνω από 20 πριν αρχίσει η πτώση. Στην συνέχεια το σκορ δείχνει μια τάση προς σταθεροποίηση, αν και με συνεχιζόμενη μεταβλητότητα.



Εικόνα 20 Το σκορ του πράκτορα Duel DQN.

Στην Εικόνα 21 παρουσιάζεται η ταχύτητα εκπαίδευσης του πράκτορα Dueling DQN σε όρους βήματα/δευτερόλεπτο κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Το διάγραμμα αποτυπώνει την απόδοση του συστήματος με την πάροδο του χρόνου, δίνοντας μια εικόνα για την αποδοτικότητα της διαδικασίας εκπαίδευσης.

Στην αρχική περίοδο της εκπαίδευσης παρατηρείται υψηλή ταχύτητα εκπαίδευσης, φτάνοντας μέχρι και πάνω από 800 βήματα/δευτερόλεπτο. Μετά την αρχική περίοδο, η ταχύτητα σταθεροποιείται γύρω στα 200 βήματα/δευτερόλεπτο. Τα ίδια σχόλια για την ταχύτητα εκπαίδευσης των προηγούμενων πρακτόρων ισχύουν και για τον Dueling DQN.



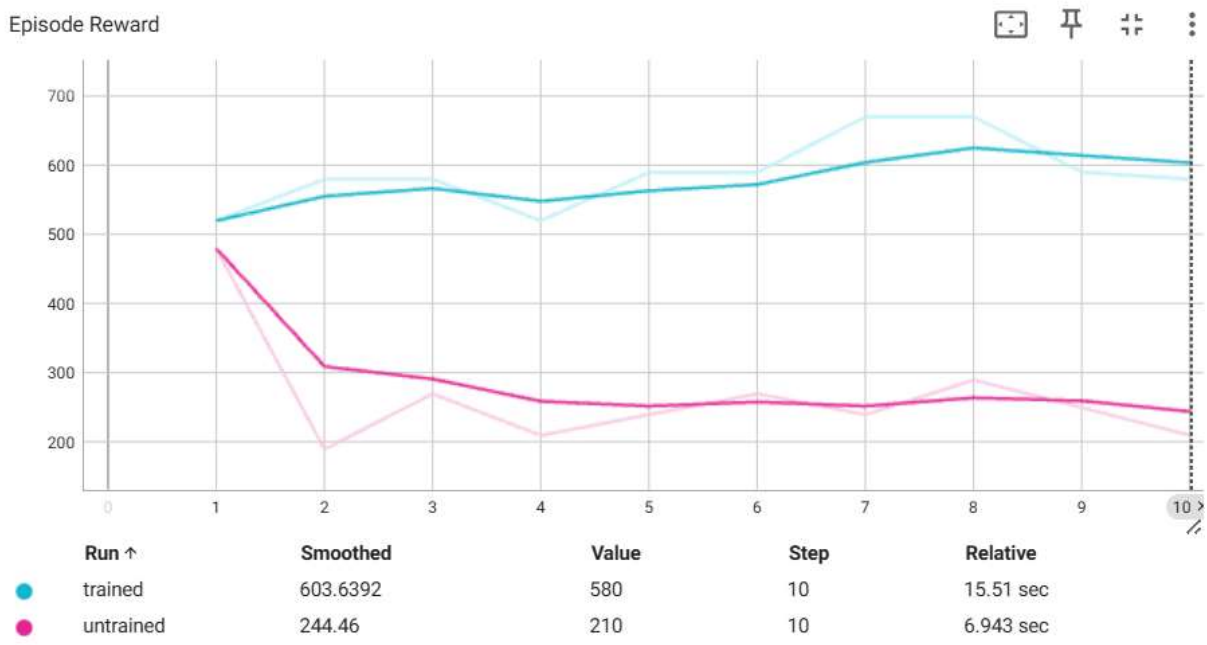
Εικόνα 21 Η ταχύτητα εκπαίδευσής του πράκτορα Duel DQN.

Στην εικόνα 22 έχουμε τη σύγκριση των ανταμοιβών που επιτυγχάνονται από έναν εκπαιδευμένο και έναν ανεκπαίδευτο πράκτορα DQN κατά τα πρώτα 10 επεισόδια. Το διάγραμμα αποτυπώνει την επίδοση των δύο πρακτόρων σε όρους ανταμοιβής ανά επεισόδιο.

Το εκπαιδευμένο μοντέλο ξεκινά με υψηλότερη ανταμοιβή (περίπου 600) σε σύγκριση με το ανεκπαίδευτο μοντέλο (περίπου 400), οπότε το εκπαιδευμένο μοντέλο έχει ήδη μάθει αποτελεσματικές στρατηγικές από την εκπαίδευση.

Το εκπαιδευμένο μοντέλο δείχνει σταθερή βελτίωση και υψηλότερη απόδοση, παραμένοντας πάνω από τα 500. Το ανεκπαίδευτο μοντέλο παραμένει χαμηλότερο στις ανταμοιβές, με μικρές αυξομειώσεις, υποδεικνύοντας αδυναμία να αναπτύξει αποτελεσματικές στρατηγικές μέσα σε αυτά τα επεισόδια.

Η σύγκριση των ανταμοιβών για τα πρώτα 10 επεισόδια ανάμεσα στον εκπαιδευμένο και τον ανεκπαίδευτο πράκτορα DQN δείχνει ξεκάθαρα την υπεροχή του εκπαιδευμένου μοντέλου. Το εκπαιδευμένο μοντέλο ξεκινά με υψηλότερη ανταμοιβή και διατηρεί σταθερά καλύτερες επιδόσεις σε σχέση με το ανεκπαίδευτο μοντέλο καθ' όλη τη διάρκεια των πρώτων 10 επεισοδίων.

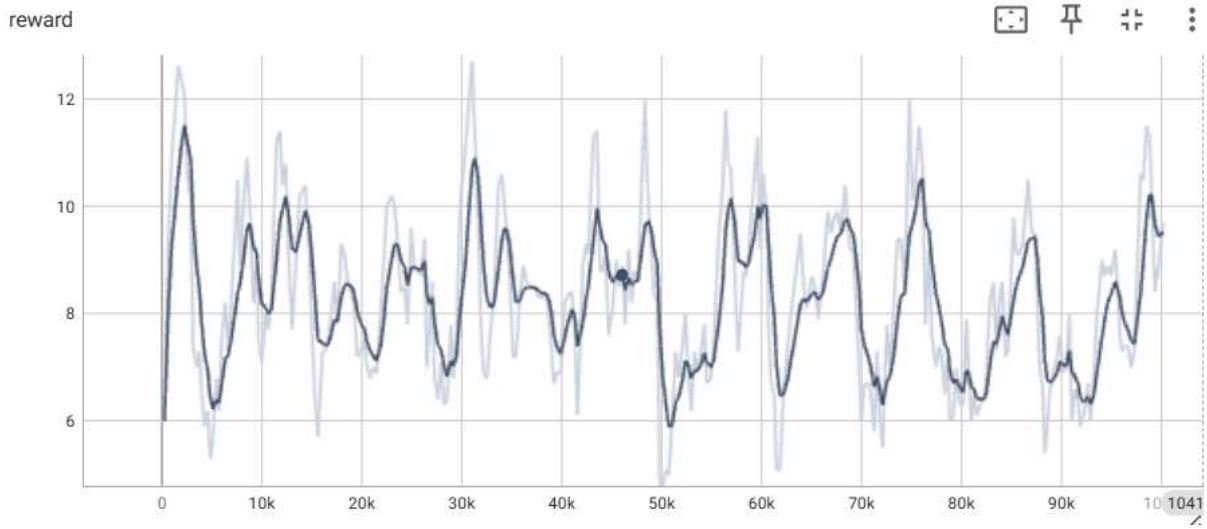


Εικόνα 22 Σύγκριση του εκπαιδευμένου πράκτορα duel DQN με έναν ανεκπαίδευτο.

5.4 Πράκτορας Noisy DQN

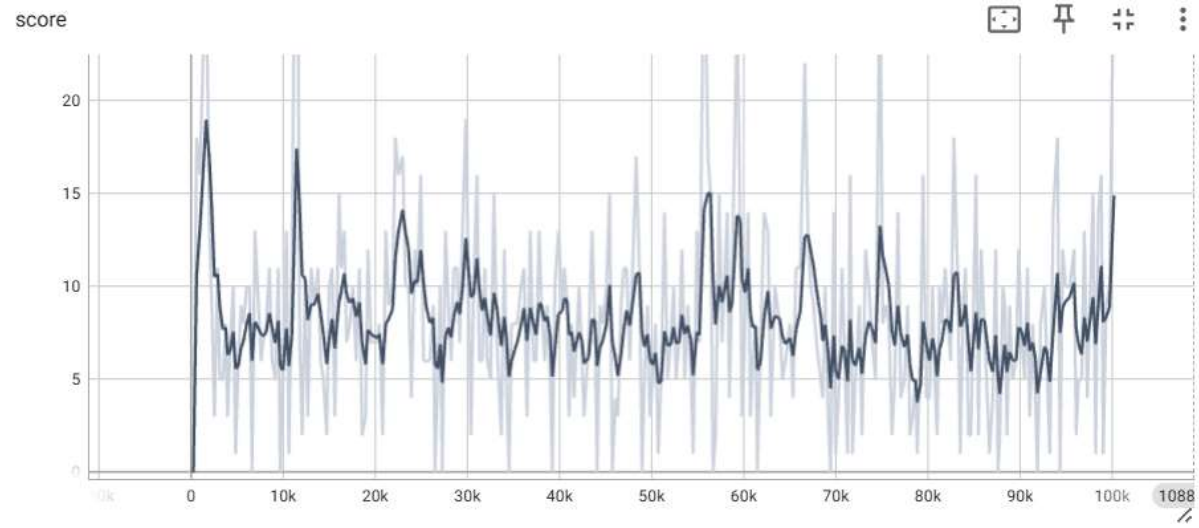
Στην Εικόνα 23 παρουσιάζεται η μέση ανταμοιβή που επιτυγχάνεται από τον πράκτορα Noisy DQN κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Όπως και με τους πράκτορες DQN και Duel DQN, η μεταβλητότητα στην απόδοση είναι εμφανής και στον πράκτορα Noisy DQN. Αυτό υποδηλώνει ότι η σταθερότητα της μάθησης παραμένει πρόκληση για όλους τους πράκτορες. Οι διακυμάνσεις στην απόδοση του πράκτορα Noisy DQN μπορεί να είναι πιο έντονες λόγω του θορύβου που εισάγεται για τη διερεύνηση.

Η αρχική ανοδική τάση στον πράκτορα Noisy DQN είναι παρόμοια με αυτή του πράκτορα Duel DQN, αλλά ακολουθείται από μια πιο απότομη πτώση.
Ο πράκτορας Noisy DQN φαίνεται να έχει περισσότερες και συχνότερες διακυμάνσεις από τον πράκτορα DQN, κάτι που μπορεί να αποδοθεί στον επιπλέον θόρυβο που χρησιμοποιείται για διερεύνηση.



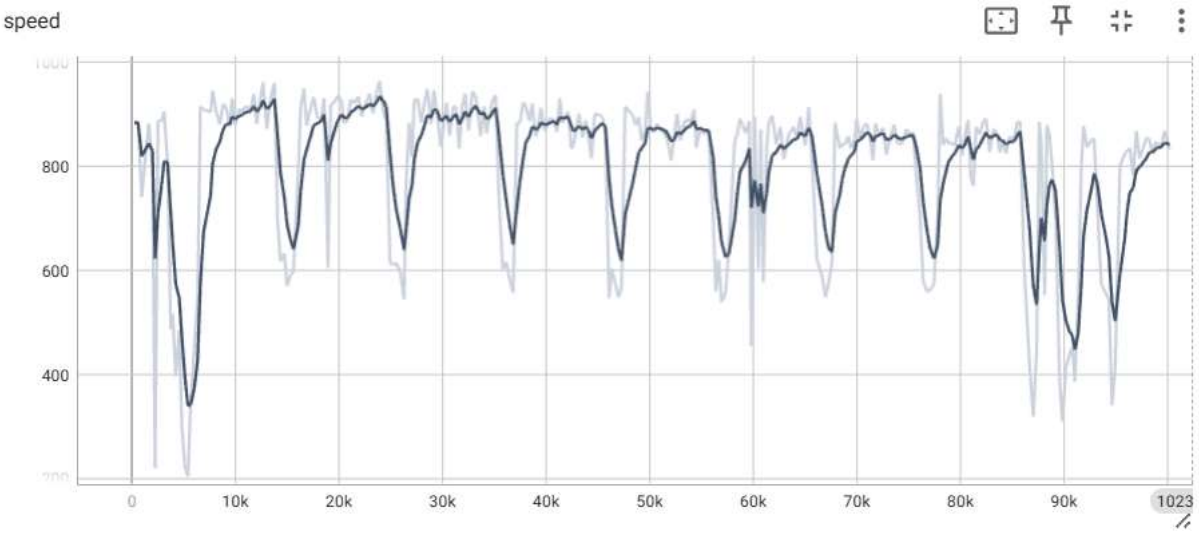
Εικόνα 23 Η μέση ανταμοιβή για κάθε δύο επεισόδια για τον πράκτορα Noisy DQN.

Στην Εικόνα 24 παρουσιάζεται το σκορ που επιτυγχάνεται από τον πράκτορα Noisy DQN κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Στην αρχική περίοδο της εκπαίδευσης παρατηρείται απότομη αύξηση στο σκορ, φτάνοντας μέχρι και πάνω από 20 πριν αρχίσει η πτώση. Αυτή η γρήγορη άνοδος είναι χαρακτηριστική καθώς ο πράκτορας προσαρμόζεται γρήγορα στο περιβάλλον και μαθαίνει βασικές στρατηγικές. Παρά τη συνεχιζόμενη μεταβλητότητα, το Noisy DQN καταφέρνει να διατηρεί μια πιο σταθερή μέση ανταμοιβή με υψηλότερες αιχμές, υποδεικνύοντας ότι η εισαγωγή θορύβου στις γραμμικές στρώσεις του νευρωνικού δικτύου είναι αποτελεσματική στην ενίσχυση της απόδοσης και της σταθερότητας του πράκτορα κατά τη διάρκεια της ενισχυτικής μάθησης.



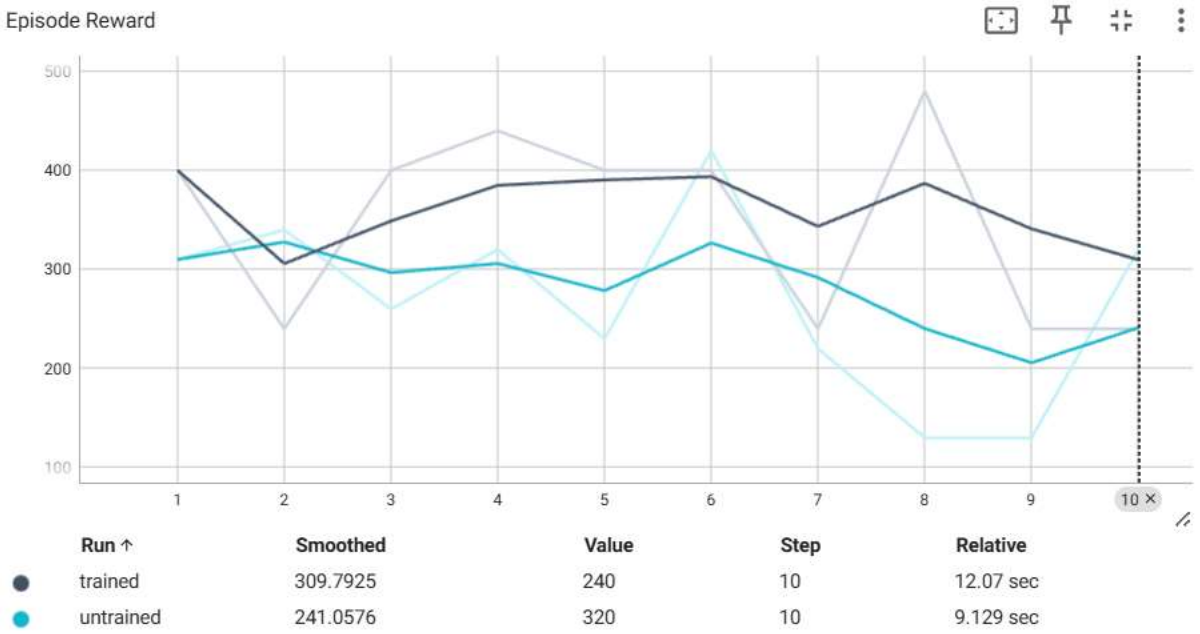
Εικόνα 24 Το σκορ του πράκτορα Noisy DQN.

Στην Εικόνα 25 παρουσιάζεται η ταχύτητα εκπαίδευσης του πράκτορα Noisy DQN σε όρους βήματα/δευτερόλεπτο κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Στην αρχική περίοδο της εκπαίδευσης παρατηρείται υψηλή ταχύτητα εκπαίδευσης, φτάνοντας μέχρι και πάνω από 800 βήματα/δευτερόλεπτο. Στην συνέχεια παρατηρείται μια σταθεροποίηση της ταχύτητας γύρω από τα 800 βήματα/δευτερόλεπτο με περιοδικές πτώσεις σε χαμηλότερα επίπεδα.
Υπάρχουν αρκετές διακυμάνσεις, με την ταχύτητα να πέφτει σε χαμηλότερα επίπεδα και να επανέρχεται. Το Noisy DQN καταφέρνει να διατηρεί υψηλότερη ταχύτητα εκπαίδευσης, υποδεικνύοντας την αποτελεσματικότητα της χρήσης θορύβου για την διατήρηση της ταχύτητας εξερεύνησης.



Εικόνα 25 Η ταχύτητα εκπαίδευσής του πράκτορα Noisy DQN.

Στην Εικόνα 26 έχουμε την σύγκριση του εκπαιδευμένου πράκτορα με έναν ανεκπαίδευτο. Όπως και στα άλλα μοντέλα, ο εκπαιδευμένος πράκτορας υπερτερεί του ανεκπαίδευτου.

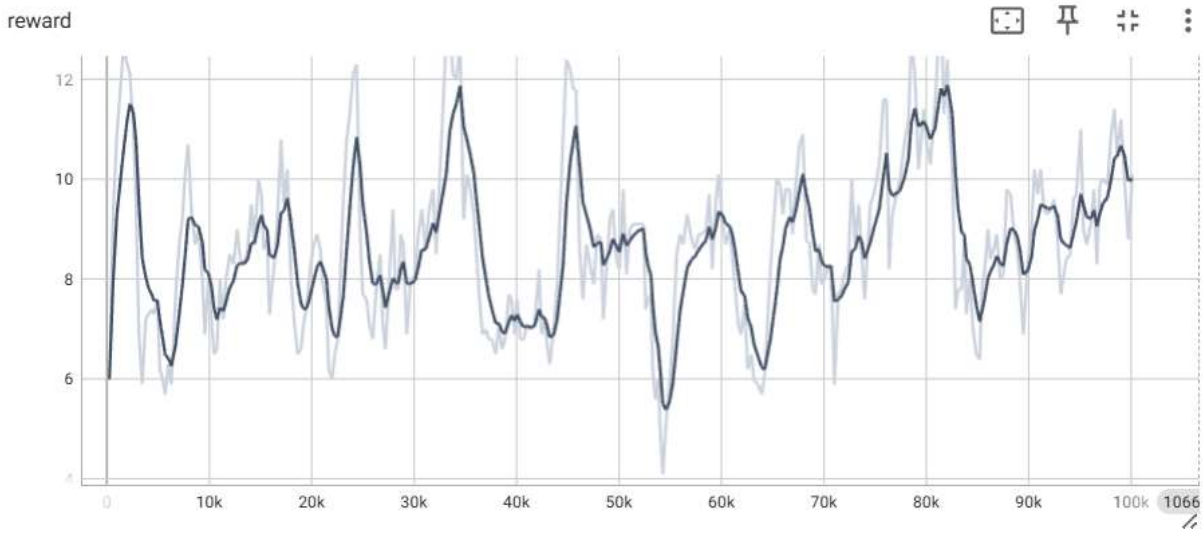


Εικόνα 26 Σύγκριση του εκπαιδευμένου πράκτορα noisy DQN με έναν ανεκπαίδευτο.

5.5 Πράκτορας double DQN με prioritized experience buffer

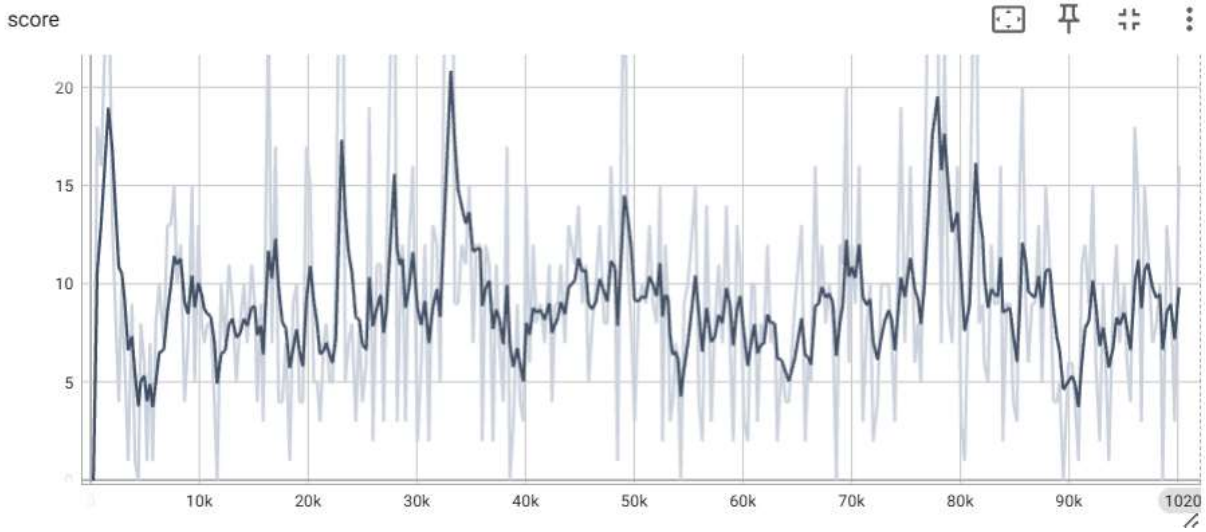
Στην Εικόνα 27 παρουσιάζεται η μέση ανταμοιβή που επιτυγχάνεται από τον πράκτορα Double DQN με Prioritized Experience Buffer κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Στην αρχική περίοδο της εκπαίδευσης παρατηρείται απότομη αύξηση στη μέση ανταμοιβή, φτάνοντας μέχρι και πάνω από 12 πριν αρχίσει η πτώση.

Η χρήση του prioritized experience buffer μπορεί να συμβάλλει στην καλύτερη αξιοποίηση των σημαντικών εμπειριών και είναι αποτελεσματικός στην ενίσχυση της απόδοσης του πράκτορα κατά τη διάρκεια της ενισχυτικής μάθησης..



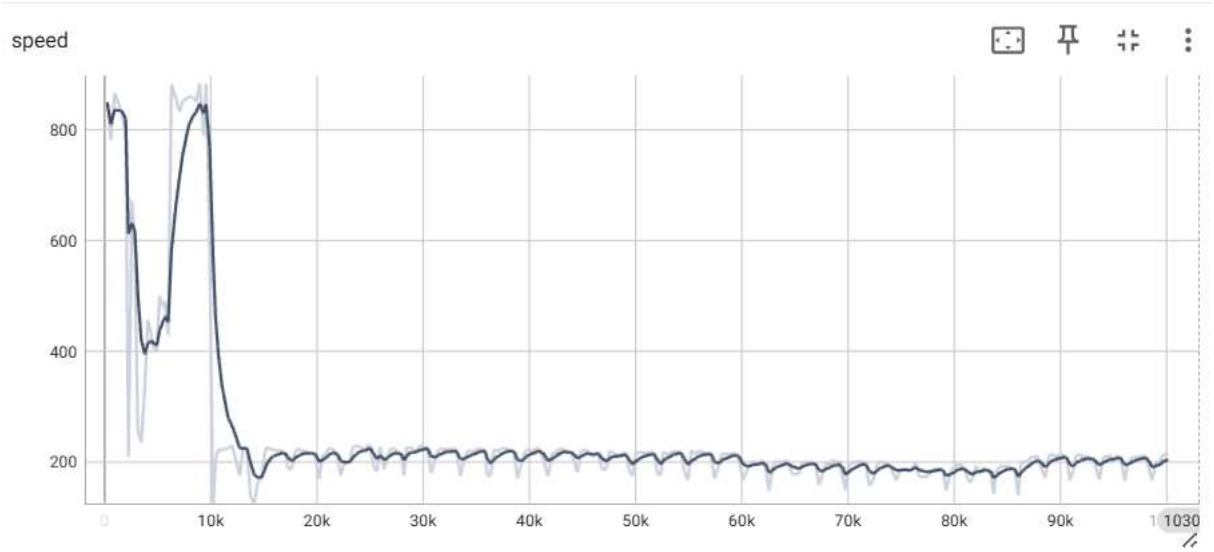
Εικόνα 27 Η μέση ανταμοιβή για κάθε δύο επεισόδια για τον πράκτορα double DQN με prioritized experience buffer.

Στην Εικόνα 28 παρουσιάζεται το σκορ που επιτυγχάνεται από τον πράκτορα Double DQN με Prioritized Experience Buffer κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Στην αρχική περίοδο της εκπαίδευσης παρατηρείται απότομη αύξηση στο σκορ, φτάνοντας μέχρι και πάνω από 20 πριν αρχίσει η πτώση. Το σκορ δείχνει μια τάση προς σταθεροποίηση, αν και με συνεχιζόμενη μεταβλητότητα. Παρατηρούνται αιχμές προς το τέλος της εκπαίδευσης που φτάνουν σε υψηλές τιμές.



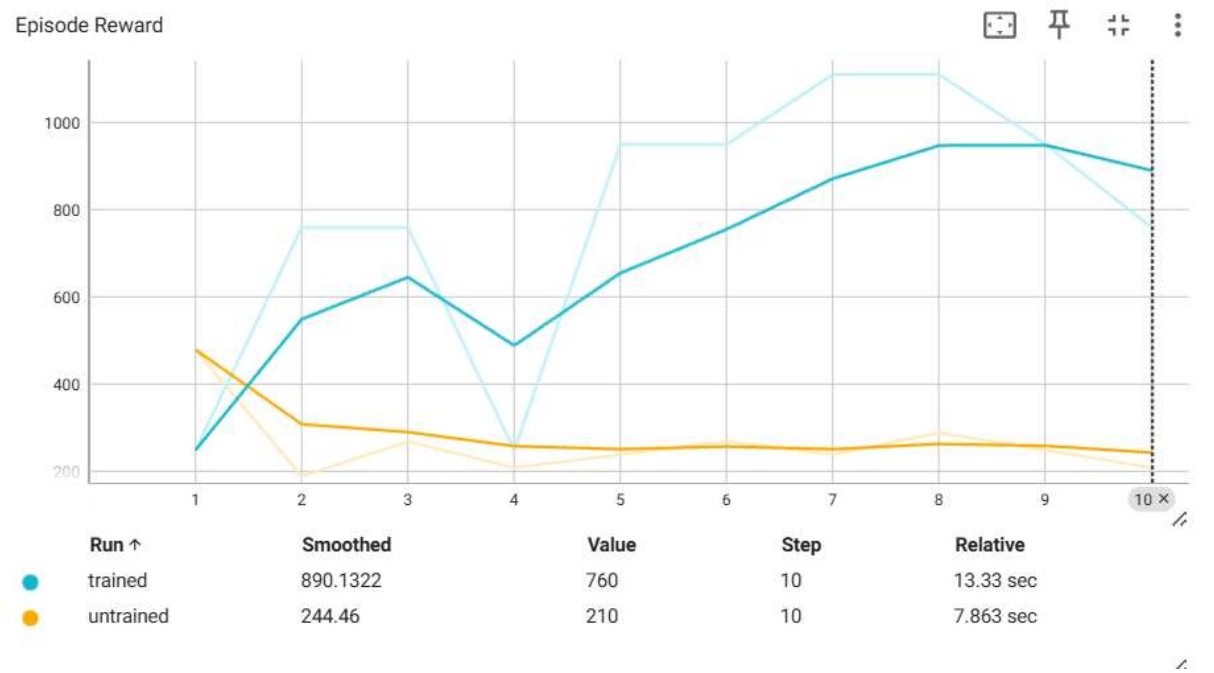
Εικόνα 28. Το σκορ του πράκτορα DQN με prioritized experience buffer.

Στην Εικόνα 29 έχουμε την ταχύτητα εκπαίδευσης του πράκτορα Double DQN με Prioritized Experience Buffer σε όρους βήματα/δευτερόλεπτο κατά τη διάρκεια της εκπαίδευσης σε διάφορα επεισόδια. Στην αρχική περίοδο της εκπαίδευσης παρατηρείται υψηλή ταχύτητα εκπαίδευσης, φτάνοντας μέχρι και πάνω από 800 βήματα/δευτερόλεπτο. Στην συνέχεια έχουμε μια σταθεροποίηση της ταχύτητας γύρω από τα 200 βήματα/δευτερόλεπτο



Εικόνα 29 Η ταχύτητα εκπαίδευσής του πράκτορα DQN με prioritized replay buffer.

Η παρακάτω εικόνα παρουσιάζει τη σύγκριση των ανταμοιβών που επιτυγχάνονται από έναν εκπαιδευμένο και έναν ανεκπαίδευτο πράκτορα DQN κατά τα πρώτα 10 επεισόδια. Η σύγκριση των ανταμοιβών για τα πρώτα 10 επεισόδια ανάμεσα στον εκπαιδευμένο και τον ανεκπαίδευτο πράκτορα DQN δείχνει ξεκάθαρα την υπεροχή του εκπαιδευμένου μοντέλου.



Εικόνα 30 Σύγκριση του εκπαιδευμένου πράκτορα DQN με prioritized replay buffer με έναν ανεκπαίδευτο.

6 Συζήτηση

Στο παρόν κεφάλαιο, παρουσιάζεται μια εκτενής ανάλυση των αποδόσεων και της ταχύτητας εκπαίδευσης διαφόρων παραλλαγών του αλγορίθμου Deep Q-Network (DQN) στο περιβάλλον MsPacmanNoFrameskip-v4. Οι παραλλαγές που εξετάστηκαν περιλαμβάνουν τον βασικό DQN, τον DQN με batch normalization, τον Duel DQN, τον Noisy DQN και τον DQN με prioritized experience buffer. Η ανάλυση βασίζεται σε δεδομένα από τα πειράματα που πραγματοποιήθηκαν και παρουσιάζονται μέσω διαγραμμάτων.

6.1 Απόδοση των Πρακτόρων

Η απόδοση κάθε πράκτορα αξιολογήθηκε με βάση το μέσο σκορ που επιτυγχάνει κατά τη διάρκεια της εκπαίδευσης.

Βασικός DQN

Η απόδοση του βασικού DQN παρουσιάζει σημαντική διακύμανση στην αρχική φάση των πρώτων 10k επεισοδίων, με αιχμές που φτάνουν στους 20 πόντους. Στα επόμενα επεισόδια, παρατηρείται σταθεροποίηση γύρω από το σκορ των 10 πόντων, με μικρότερες διακυμάνσεις. Μετά τα 60k επεισόδια, η απόδοση δείχνει τάση βελτίωσης, φτάνοντας κοντά στους 15 πόντους μέχρι τα 100k επεισόδια, με περιορισμένες διακυμάνσεις.

DQN με Batch Normalization

Η απόδοση του DQN με Batch Normalization εμφανίζει επίσης υψηλή διακύμανση στην αρχική φάση, με αιχμές στους 20 πόντους. Στα πρώτα 10k επεισόδια, η απόδοση σταθεροποιείται γύρω στους 10 πόντους, με ηπιότερες διακυμάνσεις. Μετά τα 60k επεισόδια, η βελτίωση είναι πιο ομαλή και σταθερή, φτάνοντας κοντά στους 15 πόντους.

Dueling DQN

Η απόδοση του Dueling DQN παρουσιάζει μεγαλύτερη αρχική διακύμανση, με αιχμές στους 25 πόντους. Στα πρώτα 10k επεισόδια, η απόδοση μειώνεται και σταθεροποιείται γύρω στους 10 πόντους, με εμφανείς διακυμάνσεις. Μετά τα 60k επεισόδια, η βελτίωση είναι σταδιακή, φτάνοντας κοντά στους 15 πόντους.

Noisy DQN

Η απόδοση του Noisy DQN εμφανίζει υψηλή αρχική διακύμανση, με αιχμές στους 15 πόντους. Στα πρώτα 10k επεισόδια, η απόδοση μειώνεται και σταθεροποιείται γύρω στους 10 πόντους, με εμφανείς διακυμάνσεις. Μετά τα 60k επεισόδια, η απόδοση βελτιώνεται σταδιακά, φτάνοντας κοντά στους 12 πόντους.

Double DQN με Prioritized Experience Buffer

Η απόδοση του Double DQN με Prioritized Experience Buffer εμφανίζει υψηλή αρχική διακύμανση, με αιχμές στους 25 πόντους. Στα πρώτα 10k επεισόδια, η απόδοση μειώνεται και σταθεροποιείται γύρω στους 10 πόντους, με εμφανείς διακυμάνσεις. Μετά τα 60k επεισόδια, η απόδοση βελτιώνεται σταδιακά, φτάνοντας κοντά στους 15 πόντους.

6.2 Ταχύτητα Εκπαίδευσης των Πρακτόρων

Η ταχύτητα εκπαίδευσης κάθε πράκτορα αξιολογήθηκε με βάση τον αριθμό των βημάτων που εκτελούνται ανά δευτερόλεπτο κατά τη διάρκεια της εκπαίδευσης.

Βασικός DQN

Η ταχύτητα εκπαίδευσης του βασικού DQN ξεκινά από περίπου 800 βήματα ανά δευτερόλεπτο, μειώνεται σταδιακά στα 200 βήματα ανά δευτερόλεπτο μετά τα πρώτα 10k επεισόδια, και παραμένει σχετικά σταθερή σε αυτό το επίπεδο για το υπόλοιπο της εκπαίδευσης.

DQN με Batch Normalization

Η ταχύτητα εκπαίδευσης του DQN με Batch Normalization ξεκινά από περίπου 800 βήματα ανά δευτερόλεπτο, μειώνεται σταδιακά στα 200 βήματα ανά δευτερόλεπτο στα πρώτα 10k επεισόδια, και παραμένει σταθερή σε αυτό το επίπεδο για το υπόλοιπο της εκπαίδευσης.

Dueling DQN

Η ταχύτητα εκπαίδευσης του Dueling DQN ξεκινά από περίπου 800 βήματα ανά δευτερόλεπτο, μειώνεται γρήγορα στα 200 βήματα ανά δευτερόλεπτο στα πρώτα 10k επεισόδια, και παραμένει σταθερή για το υπόλοιπο της εκπαίδευσης.

Noisy DQN

Η ταχύτητα εκπαίδευσης του Noisy DQN ξεκινά από περίπου 800 βήματα ανά δευτερόλεπτο, παρουσιάζει περιοδικές πτώσεις κάτω από τα 400 βήματα ανά δευτερόλεπτο, και σταθεροποιείται γύρω στα 800 βήματα ανά δευτερόλεπτο για το υπόλοιπο της εκπαίδευσης.

Double DQN με Prioritized Experience Buffer

Η ταχύτητα εκπαίδευσης του Double DQN με Prioritized Experience Buffer ξεκινά από περίπου 800 βήματα ανά δευτερόλεπτο, μειώνεται σταδιακά στα 200 βήματα ανά δευτερόλεπτο μετά τα πρώτα 10k επεισόδια, και παραμένει σταθερή σε αυτό το επίπεδο για το υπόλοιπο της εκπαίδευσης.

6.3 Συγκριτική Ανάλυση

Απόδοση

Οι πράκτορες που χρησιμοποίησαν Batch Normalization και Prioritized Experience Buffer παρουσίασαν πιο σταθερές αποδόσεις με ηπιότερες διακυμάνσεις κατά τη διάρκεια της εκπαίδευσης. Συγκεκριμένα, ο DQN με Batch Normalization πέτυχε την υψηλότερη σταθερότητα και ομαλότητα στη βελτίωση της απόδοσης, ενώ ο Double DQN με Prioritized Experience Buffer παρουσίασε υψηλή αρχική διακύμανση, αλλά σταδιακή βελτίωση της απόδοσης.

Ο Dueling DQN, παρά την υψηλή αρχική διακύμανση και τις αιχμές της 25 πόντους, παρουσίασε μια

πιο αργή και σταδιακή βελτίωση στη συνέχεια. Ο Noisy DQN εμφάνισε ενδιαφέρουσες τάσεις, με υψηλή αρχική διακύμανση και σταθεροποίηση γύρω της 10 πόντους, αλλά με συνεχείς διακυμάνσεις στην τελική φάση της εκπαίδευσης.

Ταχύτητα Εκπαίδευσης

Ο DQN με Batch Normalization εμφάνισε σταθερή ταχύτητα εκπαίδευσης γύρω στα 200 βήματα ανά δευτερόλεπτο μετά την αρχική περίοδο, ενώ ο Noisy DQN παρουσίασε περιοδικές πτώσεις και αυξήσεις στην ταχύτητα εκπαίδευσης. Ο βασικός DQN και ο Double DQN με Prioritized Experience Buffer παρουσίασαν παρόμοια συμπεριφορά στην ταχύτητα εκπαίδευσης, με αρχικά υψηλή ταχύτητα που μειώθηκε σταθερά σε περίπου 200 βήματα ανά δευτερόλεπτο.

6.4 Συμπεράσματα για την Ταχύτητα Εκπαίδευσης και την Απόδοση

Η χρήση Batch Normalization στον DQN φαίνεται να διατηρεί υψηλότερη σταθερότητα στην απόδοση, ενώ ο Noisy DQN παρουσιάζει μεγάλη διακύμανση στην ταχύτητα και την απόδοση. Ο Double DQN με Prioritized Experience Buffer παρουσιάζει χαμηλή αρχική ταχύτητα, αλλά σταδιακή βελτίωση στην πορεία της εκπαίδευσης, ενώ η απόδοσή του είναι ανταγωνιστική με αυτή του βασικού DQN.

7 Συμπεράσματα

Η παρούσα διπλωματική εργασία εξετάζει τη χρήση της ενισχυτικής μάθησης για την εκπαίδευση πρακτόρων σε περιβάλλοντα ηλεκτρονικών παιχνιδιών. Οι δοκιμές που πραγματοποιήθηκαν περιλαμβάνουν διάφορες παραλλαγές του αλγορίθμου DQN (Deep Q-Network) και η απόδοσή τους συγκρίνεται βάσει της μέσης ανταμοιβής, του σκορ και της ταχύτητας εκπαίδευσης.

7.1 Ανακεφαλαίωση Αποτελεσμάτων

Πράκτορας DQN

Μέση Ανταμοιβή: Ο πράκτορας DQN παρουσίασε σταθερή βελτίωση της μέσης ανταμοιβής με την πρόοδο των επεισοδίων.

Σκορ: Το σκορ του πράκτορα DQN αυξήθηκε προοδευτικά, επιδεικνύοντας την ικανότητα του πράκτορα να μαθαίνει από το περιβάλλον.

Ταχύτητα Εκπαίδευσης: Η ταχύτητα εκπαίδευσης ήταν ικανοποιητική, με τον πράκτορα να επιτυγχάνει σταθερή βελτίωση χωρίς μεγάλες διακυμάνσεις.

Πράκτορας DQN με Batch Normalization

Μέση Ανταμοιβή: Η ενσωμάτωση της batch normalization βελτίωσε τη μέση ανταμοιβή, καθιστώντας την εκπαίδευση πιο σταθερή.

Σκορ: Η συνολική απόδοση του πράκτορα παρουσίασε βελτίωση, με υψηλότερα σκορ σε σύγκριση με τον βασικό πράκτορα DQN.

Ταχύτητα Εκπαίδευσης: Η διαδικασία εκπαίδευσης ήταν ταχύτερη και πιο αποτελεσματική.

Πράκτορας Dueling DQN

Μέση Ανταμοιβή: Ο πράκτορας Dueling DQN παρουσίασε σημαντική βελτίωση στη μέση ανταμοιβή, κυρίως λόγω της διαχωρισμένης εκτίμησης αξίας και πλεονεκτήματος.

Σκορ: Η απόδοση του πράκτορα ήταν ανώτερη, με υψηλότερα σκορ σε διάφορα στάδια της εκπαίδευσης.

Ταχύτητα Εκπαίδευσης: Η ταχύτητα εκπαίδευσης βελτιώθηκε, επιτυγχάνοντας καλύτερα αποτελέσματα σε μικρότερο χρόνο.

Πράκτορας Noisy DQN

Μέση Ανταμοιβή: Ο πράκτορας Noisy DQN εμφάνισε σημαντική σταθερότητα και υψηλή μέση ανταμοιβή, αξιοποιώντας τις τυχαίες παραλλαγές στο δίκτυο.

Σκορ: Το σκορ αυξήθηκε σταθερά, καταδεικνύοντας την αποτελεσματικότητα των θορυβωδών παραμέτρων.

Ταχύτητα Εκπαίδευσης: Παρά τις τυχαίες παραλλαγές, η ταχύτητα εκπαίδευσης παρέμεινε ικανοποιητική.

Double DQN με Prioritized Experience Buffer

Μέση Ανταμοιβή: Ο πράκτορας παρουσίασε την υψηλότερη μέση ανταμοιβή, καθώς η προτεραιοποίηση εμπειριών βελτιστοποίησε τη διαδικασία εκπαίδευσης.

Σκορ: Το σκορ του πράκτορα ήταν το υψηλότερο μεταξύ όλων των παραλλαγών, υποδεικνύοντας τη βελτιωμένη απόδοση.

Ταχύτητα Εκπαίδευσης: Η ταχύτητα εκπαίδευσης ήταν ιδιαίτερα αυξημένη, με ταχύτερη σύγκλιση και βελτιστοποίηση των παραμέτρων.

7.2 Θεωρητική Ανάλυση

Η θεωρία της ενισχυτικής μάθησης, της αναλύθηκε στην εργασία, περιλαμβάνει της βασικές αρχές της πολιτικής, της συνάρτησης ανταμοιβής, της συνάρτησης αξίας και του μοντέλου περιβάλλοντος. Η υλοποίηση των παραπάνω αρχών στα διάφορα μοντέλα DQN και οι επεκτάσεις της δείχνουν πώς η θεωρία εφαρμόζεται στην πράξη και πώς οι βελτιώσεις της μεθοδολογίες μπορούν να οδηγήσουν σε αυξημένες επιδόσεις και καλύτερη απόδοση των πρακτόρων.

Η χρήση τεχνικών όπως το batch normalization, το duel DQN και οι θορυβώδεις παραλλαγές επιβεβαιώνουν την ανάγκη για σταθερότητα και αποδοτικότητα στην εκπαίδευση πρακτόρων, ενώ η προτεραιοποίηση εμπειριών (prioritized experience buffer) αναδεικνύει τη σημασία της στρατηγικής επιλογής δειγμάτων για τη βελτίωση της εκμάθησης.

7.3 Συμπεράσματα για την Ταχύτητα Εκπαίδευσης και την Απόδοση

Σε γενικές γραμμές, τα αποτελέσματα αυτής της μελέτης καταδεικνύουν τη σημασία των προσαρμογών στις βασικές αρχές της ενισχυτικής μάθησης για τη βελτίωση των επιδόσεων σε σύνθετα και δυναμικά περιβάλλοντα όπως τα ηλεκτρονικά παιχνίδια. Η θεωρητική κατανόηση των θεμελιωδών εννοιών επιτρέπει την ανάπτυξη καινοτόμων αλγορίθμων που μπορούν να προσαρμοστούν και να βελτιστοποιηθούν για διάφορες εφαρμογές.

7.4 Ανακεφαλαίωση

Η χρήση της ενισχυτικής μάθησης για την εκπαίδευση πρακτόρων σε ηλεκτρονικά παιχνίδια αποδεικνύεται αποτελεσματική, με τους πράκτορες να παρουσιάζουν βελτιωμένη απόδοση μέσω των διαφόρων τεχνικών και μεθοδολογιών που εφαρμόστηκαν. Οι πρακτόρες που χρησιμοποίησαν batch normalization και prioritized experience buffer παρουσίασαν πιο σταθερές αποδόσεις με ηπιότερες διακυμάνσεις κατά τη διάρκεια της εκπαίδευσης. Συγκεκριμένα, ο DQN με batch normalization πέτυχε την υψηλότερη σταθερότητα και ομαλότητα στη βελτίωση της απόδοσης, ενώ ο Double DQN με Prioritized Experience Buffer παρουσίασε υψηλή αρχική διακύμανση, αλλά σταδιακή βελτίωση της απόδοσης.

Βιβλιογραφία

[1]. Alam, A. (2022, April). A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 69-74). IEEE.

[2]. Almahdi, S., & Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, *87*, 267-279.

[3]. Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, *47*, 253-279.

[4]. Bertens, P., Guitart, A., Chen, P. P., & Perianez, A. (2018, August). A machine-learning item recommendation system for video games. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 1-4). IEEE.

[5]. Bertolini, M., Mezzogori, D., Neroni, M., & Zammori, F. (2021). Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, *175*, 114820.

[6]. Busoniu, L., Babuska, R., De Schutter, B., & Ernst, D. (2017). *Reinforcement learning and dynamic programming using function approximators*. CRC press.

[7]. de Almeida Rocha, D., & Duarte, J. C. (2019, October). Simulating human behaviour in games using machine learning. In *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)* (pp. 163-172). IEEE.

[8]. Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP journal*, *1*(1), 30.

[9]. Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

[10]. Florensa, C., Held, D., Geng, X., & Abbeel, P. (2018, July). Automatic goal generation for reinforcement learning agents. In *International conference on machine learning* (pp. 1515-1528). PMLR.

[11]. Frutos-Pascual, M., & Zapirain, B. G. (2015). Review of the use of AI techniques in serious games: Decision making and machine learning. *IEEE Transactions on Computational Intelligence and AI in Games*, *9*(2), 133-152.

[12]. Gülü, M., Yagin, F. H., Gocer, I., Yapici, H., Ayyildiz, E., Clemente, F. M., ... & Nobari, H. (2023). Exploring obesity, physical activity, and digital game addiction levels among

adolescents: A study on machine learning-based prediction of digital game addiction. *Frontiers in Psychology*, *14*, 1097145.

[13]. Hitar-Garcia, J. A., Moran-Fernandez, L., & Bolon-Canedo, V. (2022). Machine learning methods for predicting league of legends game outcome. *IEEE Transactions on Games*.

[14]. Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., ... & Michalewski, H. (2019). Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*.

[15]. Lample, G., & Chaplot, D. S. (2017, February). Playing FPS games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

[16]. Lanctot, M., Lockhart, E., Lespiau, J. B., Zambaldi, V., Upadhyay, S., Pérolat, J., ... & Ryan-Davis, J. (2019). OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*.

[17]. Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., ... & Graepel, T. (2017). A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, *30*.

[18]. Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

[19]. Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

[20]. Liu, F., Viano, L., & Cevher, V. (2022). Understanding deep neural function approximation in reinforcement learning via $\epsilon $-greedy exploration. *Advances in Neural Information Processing Systems*, *35*, 5093-5108.

[21]. Mania, H., Guy, A., & Recht, B. (2018). Simple random search of static linear policies is competitive for reinforcement learning. *Advances in neural information processing systems*, *31*.

[22]. Martínez, D., Alenya, G., & Torras, C. (2017). Relational reinforcement learning with guided demonstrations. *Artificial Intelligence*, *247*, 295-312.

[23]. Millington, I. (2019). *AI for Games*. CRC Press.

[24]. Pateria, S., Subagdja, B., Tan, A. H., & Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, *54*(5), 1-35.

[25]. Petter, E. A., Gershman, S. J., & Meck, W. H. (2018). Integrating models of interval timing and reinforcement learning. *Trends in cognitive sciences*, *22*(10), 911-922.

[26]. Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229.

[27]. Shao, K., Tang, Z., Zhu, Y., Li, N., & Zhao, D. (2019). A survey of deep reinforcement learning in video games. *arXiv preprint arXiv:1912.10944*.

[28]. Summerville, A., Snodgrass, S., Guzdial, M., Holmgård, C., Hoover, A. K., Isaksen, A., ... & Togelius, J. (2018). Procedural content generation via machine learning (PCGML). *IEEE Transactions on Games*, *10*(3), 257-270.

[29]. Swishchuk, A., & Vadori, N. (2017). A semi-Markovian modeling of limit order markets. *SIAM Journal on Financial Mathematics*, *8*(1), 240-273.

[30]. Szita, I. (2012). Reinforcement learning in games. In *Reinforcement Learning: State-of-the-art* (pp. 539-577). Berlin, Heidelberg: Springer Berlin Heidelberg.

[31]. van Hasselt, H., Madjiheurem, S., Hessel, M., Silver, D., Barreto, A., & Borsa, D. (2021, May). Expected eligibility traces. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 11, pp. 9997-10005).

[32]. Vodopivec, T., Samothrakis, S., & Ster, B. (2017). On monte carlo tree search and reinforcement learning. *Journal of Artificial Intelligence Research*, *60*, 881-936.

[33]. Wachi, A., & Sui, Y. (2020, November). Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning* (pp. 9797-9806). PMLR.

[34]. Yang, Y., & Wang, J. (2020). An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*.

[35]. Yannakakis, G. N., & Togelius, J. (2018). *Artificial intelligence and games* (Vol. 2, pp. 2475-1502). New York: Springer.

[36]. Zhang, L., Li, J., Shi, H., & Hwang, K. S. (2022). Multi-agent reinforcement learning by the actor-critic model with an attention interface. *Neurocomputing*, *471*, 275-284.