



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΕΛΟΠΟΝΝΗΣΟΥ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

“Ο Ρόλος της Ενισχυτικής Μάθησης σε Εφαρμογές Παιγνίων”

ΚΑΡΑΓΙΑΝΝΑΚΗΣ ΣΤΥΛΙΑΝΟΣ

ΕΠΙΒΛΕΠΩΝ: ΙΩΑΝΝΗΣ ΖΑΧΑΡΑΚΗΣ

ΠΑΤΡΑ 2025

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Ονοματεπώνυμο, Υπογραφή

2. Ονοματεπώνυμο, Υπογραφή

3. Ονοματεπώνυμο, Υπογραφή

Υπεύθυνη Δήλωση Φοιτητή

Βεβαιώνω ότι είμαι συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τη συγκεκριμένη εργασία.

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Καραγιαννάκη Στυλιανού που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης ο συγγραφέας/δημιουργός εκχωρεί στο Πανεπιστήμιο Πελοποννήσου, μη αποκλειστική άδεια χρήσης του δικαιώματος αναπαραγωγής, προσαρμογής, δημόσιου δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσής τους διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος και για όλο το χρόνο διάρκειας των δικαιωμάτων πνευματικής ιδιοκτησίας. Η ανοικτή πρόσβαση στο πλήρες κείμενο για μελέτη και ανάγνωση δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, αποθήκευση, πώληση, εμπορική χρήση, μετάδοση, διανομή, έκδοση, εκτέλεση, «μεταφόρτωση» (downloading), «ανάρτηση» (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού. Ο συγγραφέας/δημιουργός διατηρεί το σύνολο των ηθικών και περιουσιακών του δικαιωμάτων.

ΠΕΡΙΕΧΟΜΕΝΑ

1.Εισαγωγή.....	5
1.1 Τεχνητή Νοημοσύνη: Ορισμός και Σημασία.....	7
1.2 Μηχανική Μάθηση: Κατηγορίες και Μέθοδοι.....	7
1.3 Αλγόριθμοι Εκπαίδευσης Ανά Κατηγορία Προβλημάτων.....	9
1.3.1 Επιβλεπόμενη Μάθηση.....	9
1.3.2 Μη Επιβλεπόμενη Μάθηση.....	12
1.3.3 Ενισχυτική Μάθηση.....	16
2 Αλγόριθμοι και Θεμελιώδεις Έννοιες Ενισχυτικής Μάθησης.....	21
2.1 Θεμελιώδεις Αρχές Ενισχυτικής Μάθησης.....	22
2.2 Ανάλυση Βασικών Αλγορίθμων Ενισχυτικής Μάθησης.....	22
2.2.1 Q-Learning.....	22
2.2.2 Deep Q-Network (DQN).....	24
2.2.3 Policy Gradient Methods.....	27
2.2.4 Proximal Policy Optimization (PPO).....	31
2.3 Πεδία Εφαρμογής Ενισχυτικής Μάθησης.....	33
3 Αρχιτεκτονική Συστήματος και Χαρακτηριστικά Εφαρμογής.....	35
3.1 Περιγραφή Περιβάλλοντος Επίδειξης (Endless Runner).....	35
3.2 Δομή Συστήματος και Επιμέρους Συστατικά.....	36
3.3 Αλληλεπίδραση Πράκτορα – Περιβάλλοντος.....	38
3.4 Επιλογή Αλγορίθμου PPO και Ρόλος του στην Εφαρμογή.....	39
4 Περιγραφή Υλοποίησης.....	41
4.1 Τεχνικές Επιλογές (Unity, ML-Agents, Python).....	41
4.2 Εφαρμογή Αρχιτεκτονικής σε Πρακτικό Επίπεδο.....	41
4.3 Προκλήσεις και Τεχνικές Λύσεις.....	42
4.4 Διαδικασία Εκπαίδευσης.....	43
5.Μεθοδολογία Αξιολόγησης και Αποτελέσματα.....	43
5.1 Μεθοδολογία Αξιολόγησης: Μετρικές και Ο ρόλος τους.....	44
5.2 Εκτέλεση Πειραμάτων και Παραγωγή Αποτελεσμάτων.....	44
5.3 Ανάλυση και Εξαγωγή Συμπερασμάτων Αξιολόγησης.....	45
6.Συμπεράσματα και Μελλοντικές Κατευθύνσεις.....	48
6.1 Συνολική Αποτίμηση της Εργασίας.....	48
6.2 Συμπεράσματα από την Αξιολόγηση του Συστήματος.....	48
6.3 Περιορισμοί και Προβληματισμοί.....	48
6.4 Προτάσεις για Μελλοντική Βελτίωση και Επεκτάσεις.....	49

Περίληψη: Η παρούσα πτυχιακή εργασία παρουσιάζει την ανάπτυξη μιας εφαρμογής παιχνιδιών που αξιοποιεί τεχνικές ενισχυτικής μάθησης (Reinforcement Learning), με στόχο να αναδείξει τις δυνατότητες και την προσαρμοστικότητα ενός αυτόνομου πράκτορα σε ένα δυναμικό και επαναλαμβανόμενο περιβάλλον. Συγκεκριμένα, σχεδιάστηκε και υλοποιήθηκε ένα τρισδιάστατο παιχνίδι τύπου endless runner, στο οποίο ένας πράκτορας τεχνητής νοημοσύνης, εκπαιδευμένος μέσω του αλγορίθμου Proximal Policy Optimization (PPO), μαθαίνει να αποφεύγει εισερχόμενα εμπόδια εκτελώντας άλματα με ακριβή χρονισμό. Το περιβάλλον διατηρεί σταθερό επίπεδο πρόκλησης, καθώς τα εμπόδια εμφανίζονται σε τακτά χρονικά διαστήματα και κινούνται προς τον πράκτορα, ο οποίος διαθέτει μόνο μία δυνατότητα άλματος κάθε φορά. Η διαμόρφωση αυτή δίνει έμφαση στη λήψη αποφάσεων σε πραγματικό χρόνο, απαιτώντας από τον πράκτορα να αναπτύξει αποτελεσματικές στρατηγικές αντίδρασης μέσα από μια διαδικασία δοκιμής και σφάλματος. Η εργασία παρέχει εκτενές θεωρητικό υπόβαθρο, παρουσιάζοντας τις βασικές αρχές της τεχνητής νοημοσύνης, της μηχανικής μάθησης και της ενισχυτικής μάθησης, εξηγώντας τον τρόπο που αυτές οι μέθοδοι μπορούν να εφαρμοστούν σε διαδραστικά συστήματα. Ακολουθεί αναλυτική περιγραφή της αρχιτεκτονικής του συστήματος, συμπεριλαμβανομένων των τεχνικών επιλογών, των εργαλείων λογισμικού (Unity, ML-Agents) και της ενσωμάτωσης του αλγορίθμου PPO στο περιβάλλον. Τέλος, συζητούνται οι προκλήσεις που προέκυψαν κατά την ανάπτυξη και προτείνονται μελλοντικές κατευθύνσεις για ενίσχυση της εφαρμογής, όπως η αύξηση της πολυπλοκότητας του περιβάλλοντος, η δοκιμή εναλλακτικών αλγορίθμων ενισχυτικής μάθησης και η διερεύνηση περαιτέρω εφαρμογών σε συναφείς τομείς.

Λέξεις-κλειδιά: τεχνητή νοημοσύνη, ενισχυτική μάθηση, εφαρμογές παιχνιδιών, Proximal Policy Optimization (PPO), εκπαίδευση πρακτόρων, Unity, ML-Agents.

Abstract: This thesis presents the development of a gaming application utilizing reinforcement learning techniques, aiming to showcase the capabilities and adaptability of an autonomous agent in a dynamic and repetitive environment. Specifically, a 3D endless runner game was designed and implemented, where an artificial intelligence (AI) agent, trained using the Proximal Policy Optimization (PPO) algorithm, learns to avoid oncoming obstacles by executing precise jumps at the right time. The application environment maintains a constant challenge level, with obstacles appearing at regular intervals and moving toward the agent, who can only perform a single jump at any given moment. This setup emphasizes real-time decision-making, requiring the agent to learn effective timing and reaction strategies through trial and error. The thesis provides a thorough theoretical background, introducing the key principles of artificial intelligence, machine learning, and reinforcement learning, and explains how these methods can be applied to interactive systems. A detailed description of the system architecture is provided, outlining the technical choices, software tools (Unity, ML-Agents), and the integration of PPO within the environment. The implementation section covers the development process, from the construction of the 3D game world and agent perception system to the reward mechanisms and training pipeline. Finally, the thesis discusses the challenges encountered during the development process and proposes future directions for enhancement, such as extending the complexity of the environment, experimenting with alternative reinforcement learning algorithms, and exploring further applications in similar domains.

Keywords: artificial intelligence (AI), reinforcement learning (RL), gaming applications, Proximal Policy Optimization (PPO), agent training, Unity, ML-Agents.

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

Η Τεχνητή Νοημοσύνη αποτελεί έναν από τους πιο ραγδαία αναπτυσσόμενους τομείς της επιστήμης και της τεχνολογίας, με εφαρμογές που εκτείνονται από την υγειονομική περίθαλψη και την αυτόνομη οδήγηση έως τα παιχνίδια και την ψυχαγωγία. Η τεχνητή νοημοσύνη αναφέρεται στην ικανότητα των μηχανών να εκτελούν γνωστικές λειτουργίες που σχετίζονται με τον ανθρώπινο νου, όπως η αντίληψη, η λογική, η μάθηση και η επίλυση προβλημάτων (McKinsey & Company, 2024).

Ένα υποσύνολο της τεχνητής νοημοσύνης είναι η Μηχανική Μάθηση η οποία επιτρέπει στους υπολογιστές να μαθαίνουν από δεδομένα και να βελτιώνουν την απόδοσή τους χωρίς να είναι ρητά προγραμματισμένοι. Μια ιδιαίτερα ενδιαφέρουσα προσέγγιση της μηχανικής μάθησης είναι η Ενισχυτική Μάθηση, όπου ένας πράκτορας μαθαίνει μέσω αλληλεπίδρασης με το περιβάλλον του, λαμβάνοντας ενισχύσεις (ανταμοιβές ή ποινές) για τις ενέργειές του (Pakhale, D. V., & Athawale, S. V. 2024).

Οι τεχνικές μηχανικής μάθησης κατατάσσονται συνήθως σε τρεις βασικές κατηγορίες, την επιβλεπόμενη μάθηση (supervised learning), όπου ο αλγόριθμος εκπαιδεύεται με δεδομένα εισόδου και τις αντίστοιχες εξόδους, την μη επιβλεπόμενη μάθηση (unsupervised learning) που στοχεύει στην αναγνώριση προτύπων χωρίς ετικετοποιημένα δεδομένα και την ενισχυτική μάθηση (reinforcement learning), όπου ένας πράκτορας μαθαίνει μέσω αλληλεπίδρασης με το περιβάλλον και ενίσχυσης με βάση τις ενέργειές του (Russell & Norvig, 2021).

Η ενισχυτική μάθηση έχει βρει σημαντικές εφαρμογές στον τομέα των παιχνιδιών, οδηγώντας σε εντυπωσιακά επιτεύγματα στην ανάπτυξη τεχνητών πρακτόρων ικανών να μαθαίνουν στρατηγικές, να προσαρμόζονται σε δυναμικά περιβάλλοντα και να επιτυγχάνουν επιδόσεις συγκρίσιμες ή και ανώτερες από ανθρώπινους παίκτες. Ένα χαρακτηριστικό παράδειγμα είναι το AlphaStar, ένας πράκτορας ενισχυτικής μάθησης που πέτυχε επίπεδο Grandmaster στο παιχνίδι StarCraft II, χρησιμοποιώντας μεθόδους πολυπρακτορικής μάθησης και βελτιστοποίησης πολιτικών δράσης (Zhang et al., 2022).

Η παρούσα εργασία έχει ως στόχο να μελετήσει και να επιδείξει τον ρόλο της ενισχυτικής μάθησης σε εφαρμογές παιχνιδιών μέσω της ανάπτυξης ενός συστήματος επίδειξης σε περιβάλλον Unity 3D. Συγκεκριμένα, αναπτύχθηκε ένα περιβάλλον τύπου "endless runner", στο οποίο ένας πράκτορας μαθαίνει να αποφεύγει εμπόδια χρησιμοποιώντας τον αλγόριθμο Proximal Policy Optimization (PPO).

Τα βασικά ερευνητικά ερωτήματα που εξετάζονται στην εργασία είναι:

- Πώς μπορεί να υλοποιηθεί ένα περιβάλλον ενισχυτικής μάθησης σε ένα παιχνίδι τύπου "endless runner";
- Ποιος είναι ο ρόλος του αλγορίθμου PPO στη διαδικασία μάθησης του πράκτορα;

- Ποιες τεχνικές προκλήσεις και ρυθμίσεις απαιτούνται για μια επιτυχημένη εκπαίδευση πράκτορα;
- Ποια είναι τα βασικά συμπεράσματα από την εφαρμογή αυτής της μεθοδολογίας σε ένα πρακτικό σενάριο παιχνιδιού;

Τα παραπάνω ερωτήματα συνθέτουν το πλαίσιο μελέτης της παρούσας εργασίας, η οποία στοχεύει στη θεωρητική και πρακτική διερεύνηση της ενισχυτικής μάθησης μέσα από την ανάπτυξη ενός διαδραστικού συστήματος τύπου παιχνιδιού. Στη συνέχεια, παρουσιάζονται οι βασικές έννοιες που σχετίζονται με την τεχνητή νοημοσύνη, τη μηχανική μάθηση και τις επιμέρους μεθόδους εκπαίδευσης, ώστε να διαμορφωθεί ένα πλήρες υπόβαθρο για την κατανόηση των τεχνολογιών που αξιοποιούνται στο πλαίσιο της εφαρμογής.

1.1 Τεχνητή Νοημοσύνη: Ορισμός και Σημασία

Η τεχνητή νοημοσύνη έχει αποτελέσει βασικό στοιχείο της λεγόμενης «δεύτερης μηχανικής εποχής», κατά την οποία οι ψηφιακές τεχνολογίες επιτρέπουν πλέον στις μηχανές να εκτελούν γνωστικά καθήκοντα, όπως η λήψη αποφάσεων και η αναγνώριση προτύπων μετασχηματίζοντας βαθιά τομείς όπως η εργασία, η παραγωγικότητα και η οικονομία (Brynjolfsson & McAfee, 2014). Αρχικά, η τεχνητή νοημοσύνη είχε οριστεί το 1956, ως η επιστήμη και η μηχανική κατασκευής ευφύων μηχανών (McCarthy, 2007). Αργότερα, ο Wang (2008) όρισε την τεχνητή νοημοσύνη ως τον μηχανισμό που μπορεί να εκτελέσει γνωστικά καθήκοντα, όπως η μάθηση και η επίλυση προβλημάτων, χρησιμοποιώντας τεχνολογικές καινοτομίες. Πρόκειται, λοιπόν, για έναν κλάδο της επιστήμης των ηλεκτρονικών υπολογιστών, όπου δημιουργούνται τεχνολογικά συστήματα, τα οποία έχουν την ικανότητα να εκτελούν ορισμένες διεργασίες, με τον τρόπο που θα το έκανε ένας άνθρωπος, όπως η εύρεση πληροφοριών και η παραγωγή γραπτού λόγου.

1.2 Μηχανική Μάθηση: Κατηγορίες και Μέθοδοι

Η μηχανική μάθηση είναι ένας υποκλάδος της τεχνητής νοημοσύνης που επιτρέπει στους υπολογιστές να μαθαίνουν από δεδομένα και να βελτιώνονται με την εμπειρία, χωρίς να είναι ρητά προγραμματισμένοι για κάθε εργασία. Αντί να βασίζονται σε προκαθορισμένους κανόνες, τα συστήματα δημιουργούν μοντέλα από δεδομένα για να κάνουν προβλέψεις ή αποφάσεις (Shaveta, 2023)

Τα μοντέλα της μηχανικής μάθησης κατατάσσονται σε τρεις κύριες κατηγορίες, με μια τέταρτη που συνήθως εξετάζεται ξεχωριστά λόγω της διαφορετικής της φιλοσοφίας:

- **Επιβλεπόμενη Μάθηση (Supervised Learning):** Η επιβλεπόμενη μάθηση (Supervised Learning) αποτελεί μία από τις βασικές μεθόδους, όπου τα μοντέλα εκπαιδεύονται χρησιμοποιώντας σύνολα δεδομένων με ετικέτες. Δηλαδή, κάθε είσοδος συνοδεύεται από τη σωστή έξοδο, επιτρέποντας στο σύστημα να "μάθει" τη συσχέτιση μεταξύ των δύο. Αυτή η μέθοδος βρίσκει εφαρμογή σε ποικίλα πεδία, από την αναγνώριση εικόνας και ομιλίας, έως τη διάγνωση ασθενειών και την πρόβλεψη οικονομικών τάσεων (Verma et al., 2021), (Yedavalli et al., 2020). Οι πιο συνηθισμένοι αλγόριθμοι περιλαμβάνουν τις γραμμικές παλινδρομήσεις, τα δένδρα αποφάσεων, τα νευρωνικά δίκτυα και τις μηχανές υποστήριξης διανυσμάτων (SVMs), οι οποίοι μπορούν να προσαρμοστούν σε προβλήματα ταξινόμησης ή παλινδρόμησης (Nkemdilim et al., 2024), (Pironneau, 2021). Παρόλο που η επιβλεπόμενη μάθηση απαιτεί μεγάλο όγκο δεδομένων υψηλής ποιότητας, θεωρείται ιδιαίτερος αποτελεσματική όταν τα δεδομένα είναι καλά επισημασμένα και υπάρχει σαφής συσχέτιση εισόδου-εξόδου (Jung, 2018).
- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning):** Η μη επιβλεπόμενη μάθηση (Unsupervised Learning), όπου τα μοντέλα εκπαιδεύονται βάσει μη ετικετοποιημένων δεδομένων. Ο στόχος είναι η ανακάλυψη κρυφών δομών, συσχετίσεων ή προτύπων

μέσα στα δεδομένα, όπως η ομαδοποίηση (clustering) ή η μείωση διαστάσεων (Sharma & Saxena, 2021), (Sharma, 2020). Τέτοιες μέθοδοι μιμούνται την ανθρώπινη ικανότητα να μαθαίνει χωρίς άμεση καθοδήγηση και θεωρούνται πιο «βιολογικά ρεαλιστικές» σε σχέση με άλλες μορφές μάθησης (Fyfe, 2008). Εφαρμογές της μη επιβλεπόμενης μάθησης εντοπίζονται σε πολλούς τομείς, όπως στην κυβερνοασφάλεια (Kanthraj, 2016), στην ιατρική διάγνωση (Di Felice et al., 2023), αλλά και στη βελτιστοποίηση συστημάτων μέσω νευρωνικών δικτύων (Kote, 2019). Παρά τις προκλήσεις, όπως η δυσκολία στην αξιολόγηση της απόδοσης, η μη επιβλεπόμενη μάθηση είναι απαραίτητη σε περιπτώσεις όπου η συλλογή ετικετών είναι ανέφικτη ή δαπανηρή.

- **Ενισχυτική Μάθηση(Reinforcement Learning):** Η ενισχυτική μάθηση (Reinforcement Learning - RL) που βασίζεται στην ιδέα της μάθησης μέσω αλληλεπίδρασης με το περιβάλλον. Σε αντίθεση με την επιβλεπόμενη μάθηση, όπου το σύστημα διδάσκεται μέσω έτοιμων παραδειγμάτων, στην ενισχυτική μάθηση ο πράκτορας μαθαίνει μέσα από δοκιμές και λάθη, λαμβάνοντας επιβραβεύσεις ή ποινές ανάλογα με τις ενέργειες που εκτελεί (Sutton & Barto, 1998). Ο στόχος είναι η μεγιστοποίηση της αθροιστικής ανταμοιβής μέσω της σταδιακής βελτίωσης των αποφάσεων του πράκτορα. Η ενισχυτική μάθηση χρησιμοποιεί συνήθως το μαθηματικό πλαίσιο των Μαρκοβιανών Διαδικασιών Απόφασης (Markov Decision Processes), και αποτελεί σήμερα τη βάση για πολλές προηγμένες εφαρμογές, όπως η ρομποτική, τα αυτόνομα οχήματα και οι στρατηγικοί αλγόριθμοι σε παιχνίδια (Barto & Sutton, 1997), (Sandhu et al., 2024), (Ghasemi & Ebrahimi, 2024).
- **Ημι-επιβλεπόμενη Μάθηση (Semi-Supervised Learning):** Η ημι-επιβλεπόμενη μάθηση αποτελεί μια ενδιάμεση μορφή μάθησης μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης, αξιοποιώντας ένα μικρό σύνολο ετικετοποιημένων δεδομένων σε συνδυασμό με έναν μεγάλο όγκο μη ετικετοποιημένων για την εκπαίδευση μοντέλων. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε τομείς όπου η απόκτηση ετικετοποιημένων δεδομένων είναι δαπανηρή ή απαιτεί εξειδικευμένη γνώση, όπως η αναγνώριση εικόνας, η επεξεργασία φυσικής γλώσσας και η βιοϊατρική πληροφορική (Chapelle et al., 2006). Με την αξιοποίηση της εγγενώς διαθέσιμης πληροφορίας από τα μη ετικετοποιημένα δεδομένα, τα συστήματα ημι-επιβλεπόμενης μάθησης επιτυγχάνουν βελτιωμένη απόδοση σε εργασίες ταξινόμησης, περιορίζοντας παράλληλα την ανάγκη για εκτενή επισήμανση από ανθρώπους ειδικούς (van Engelen & Hoos, 2019).

Οι παραπάνω κατηγορίες μάθησης περιλαμβάνουν ευρύ φάσμα αλγορίθμων, καθένας εκ των οποίων ενδείκνυται για συγκεκριμένα είδη προβλημάτων. Στην επόμενη ενότητα, παρουσιάζονται ενδεικτικοί αλγόριθμοι κάθε κατηγορίας, μαζί με τις αρχές λειτουργίας και εφαρμογές τους.

1.3 Αλγόριθμοι Εκπαίδευσης ανά Κατηγορία Προβλημάτων

Στην παρούσα ενότητα παρουσιάζονται ενδεικτικοί αλγόριθμοι για κάθε κατηγορία μηχανικής μάθησης, όπως αυτές περιγράφηκαν στην ενότητα 1.2. Η ανάλυση επικεντρώνεται στα χαρακτηριστικά, τη λειτουργία και τις βασικές εφαρμογές τους.

1.3.1 Επιβλεπόμενη Μάθηση

Κατηγορίες Προβλημάτων

Στο πλαίσιο της επιβλεπόμενης μάθησης, οι δύο κύριες κατηγορίες προβλημάτων είναι η ταξινόμηση (classification) και η παλινδρόμηση (regression). Και στις δύο περιπτώσεις, η εκπαίδευση γίνεται με χρήση δειγμάτων τα οποία συνοδεύονται από ετικέτες που υποδεικνύουν την επιθυμητή έξοδο.

- **Ταξινόμηση (Classification):** Το μοντέλο προσπαθεί να προβλέψει τη σωστή ετικέτα για ένα δεδομένο δείγμα εισόδου. Στη διαδικασία της ταξινόμησης, το μοντέλο εκπαιδεύεται πλήρως χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης και στη συνέχεια αξιολογείται με βάση δεδομένα ελέγχου, πριν χρησιμοποιηθεί για την πρόβλεψη νέων, άγνωστων δεδομένων (Zoumana Keita, 2024).
- **Παλινδρόμηση (Regression):** Η παλινδρόμηση είναι ένας τύπος επιβλεπόμενης μηχανικής μάθησης, όπου οι αλγόριθμοι μαθαίνουν από τα δεδομένα προκειμένου να προβλέψουν συνεχείς τιμές, όπως πωλήσεις, μισθούς, βάρος ή θερμοκρασία. Για παράδειγμα, σε ένα σύνολο δεδομένων που περιλαμβάνει χαρακτηριστικά ενός σπιτιού όπως το μέγεθος του οικοπέδου, ο αριθμός υπνοδωματίων και μπάνιων, η γειτονιά κ.ά. καθώς και την τιμή πώλησης, ένας αλγόριθμος παλινδρόμησης μπορεί να εκπαιδευτεί ώστε να μάθει τη συσχέτιση μεταξύ των χαρακτηριστικών και της τιμής του ακινήτου (Moez Ali, 2022).

Κοινοί Αλγόριθμοι

Naive Bayes

Ο αλγόριθμος Naive Bayes είναι ένα μοντέλο ταξινόμησης που βασίζεται σε πιθανότητες και εκτιμά την πιθανότητα μιας ετικέτας κατηγορίας δεδομένου ενός συνόλου χαρακτηριστικών, μοντελοποιώντας την κοινή κατανομή $P(x,y)$ και εφαρμόζοντας το Θεώρημα του Bayes, με την παραδοχή της υπό συνθήκη ανεξαρτησίας μεταξύ των χαρακτηριστικών. Παρά την απλότητά του και την ισχυρή αυτή παραδοχή ανεξαρτησίας, ο Naive Bayes συχνά παρουσιάζει ανταγωνιστική απόδοση σε εφαρμογές του πραγματικού κόσμου, ιδιαίτερα όταν τα δεδομένα εκπαίδευσης είναι περιορισμένα ή υψηλής διαστασιμότητας. Έρευνες έχουν δείξει ότι συγκλίνει στον ασυμπτωτικό ρυθμό σφάλματος του πολύ ταχύτερα από διακριτικά μοντέλα, όπως η λογιστική παλινδρόμηση, γεγονός που τον καθιστά ιδιαίτερα αποτελεσματικό σε σενάρια με λίγα δεδομένα (Ng & Jordan, 2001). Αυτή η αποδοτικότητα

και η ανθεκτικότητά του είναι ιδιαίτερα χρήσιμες σε εφαρμογές όπως η ταξινόμηση κειμένων, η ανίχνευση ανεπιθύμητης αλληλογραφίας και η ιατρική διάγνωση (Hasan et al., 2023), (Aristawidya et al., 2024). Επιπλέον, βελτιώσεις όπως η Μέση Τιμή Μοντέλων Bayes (Bayesian Model Averaging) και ο Γενικευμένος Αφελής Ταξινομητής του Bayes (Generalized Naive Bayes Classifier) έχουν ενισχύσει την ευελιξία και την προγνωστική του ισχύ, καθιστώντας τον μια ισχυρή και κλιμακούμενη επιλογή για πολλά προβλήματα ταξινόμησης (Larsen, 2005).

Λογιστική Παλινδρόμηση (Logistic Regression – LR)

Η Λογιστική Παλινδρόμηση είναι ένα γραμμικό μοντέλο ταξινόμησης που ανήκει στην κατηγορία των διακριτικών μοντέλων, καθώς μοντελοποιεί απευθείας την υπό συνθήκη πιθανότητα μιας ετικέτας κατηγορίας με βάση τα χαρακτηριστικά εισόδου, με στόχο τη βελτιστοποίηση της ακρίβειας ταξινόμησης. Εκτιμά άμεσα την πιθανότητα $P(y|x)$. Έρευνες έχουν δείξει ότι η λογιστική παλινδρόμηση πετυχαίνει συνήθως χαμηλότερο ασυμπτωτικό σφάλμα σε σύγκριση με άλλα μοντέλα, όπως το Naive Bayes, ειδικά όσο αυξάνεται το μέγεθος των δεδομένων εκπαίδευσης (Ng & Jordan, 2001). Πειραματικές συγκρίσεις επιβεβαιώνουν την ανώτερη απόδοσή της σε εφαρμογές ταξινόμησης κειμένου (Hasan et al., 2023) και σε προβλέψεις σχετικές με ασθένειες (Aristawidya et al., 2024), όπου συχνά ξεπερνά τον αλγόριθμο Naive Bayes ως προς την ακρίβεια. Επιπλέον, οι εξελίξεις σε τεχνικές βελτιστοποίησης, όπως η ανάβαση κλίσης (gradient ascent) και η μέθοδος Newton-Raphson έχουν βελτιώσει σημαντικά την υπολογιστική της αποδοτικότητα και την κλιμάκωση (Bhowmik, 2015).

Δέντρα Αποφάσεων

Τα Δέντρα Αποφάσεων (Decision Trees) είναι μια ευρέως χρησιμοποιούμενη μέθοδος της επιβλεπόμενης μάθησης λόγω της απλότητας, της ερμηνευσιμότητας και της ικανότητάς τους να μοντελοποιούν πολύπλοκες διαδικασίες λήψης αποφάσεων. Ταξινομούν τα δεδομένα δημιουργώντας μία δομή δένδρου όπου οι εσωτερικοί κόμβοι αναπαριστούν αποφάσεις και τα φύλλα δηλώνουν τα αποτελέσματα. Η διαισθητική τους δομή τα καθιστά ιδιαίτερα χρήσιμα σε τομείς όπως η ιατρική διάγνωση, τα οικονομικά και η ταξινόμηση κειμένων (Bahzad et al., 2021). Τα Δέντρα Αποφάσεων είναι εύκολα στην υλοποίηση και αποδίδουν καλά με ελάχιστη προεπεξεργασία των δεδομένων. Ωστόσο, είναι επιρρεπή στην υπερπροσαρμογή (overfitting), κάτι που συχνά αντιμετωπίζεται μέσω κλαδέματος ή μεθόδων ensemble όπως τα Random Forests και οι τεχνικές boosting (Rokach & Maimon, 2005), (Coadou, 2016). Νεότερες καινοτομίες, όπως το μοντέλο Tree-in-Tree (TnT), ενισχύουν περαιτέρω τα Δέντρα Αποφάσεων, ενσωματώνοντας μικρότερα δέντρα μέσα στους κόμβους, βελτιώνοντας την ακρίβεια ενώ διατηρούν την αποδοτικότητα (Zhu & Shoaran, 2021). Αυτές

οι εξελίξεις αναδεικνύουν τη διαρκή σημασία των Δέντρων Απόφασης ως ένα βασικό και ευέλικτο εργαλείο στη μηχανική μάθηση.

Ενισχυμένα Δέντρα Αποφάσεων

Τα Ενισχυμένα Δέντρα Αποφάσεων (Boosted Decision Trees) είναι ένα σύνολο από δένδρα τα οποία εκπαιδεύονται διαδοχικά, όπου κάθε δέντρο κατασκευάζεται ώστε να διορθώνει τα σφάλματα των προηγούμενων. Οι δύο κύριες κατηγορίες των ενισχυμένων δέντρων αποφάσεων είναι τα AdaBoost και Gradient Boosting. Στο AdaBoost, κάθε δέντρο εκπαιδεύεται πάνω σε μια αναπροσαρμοσμένη εκδοχή των δεδομένων εκπαίδευσης όπου δίνονται μεγαλύτερα βάρη στα παραδείγματα που ταξινομήθηκαν λανθασμένα στις προηγούμενες φάσεις. Στο Gradient Boosting, τα δέντρα εκπαιδεύονται ώστε να προβλέψουν την κλίση της συνάρτησης απώλειας ως προς την έξοδο του μοντέλου, διορθώνοντας έτσι τα υπολειπόμενα σφάλματα των προηγούμενων δέντρων. Αυτές οι μέθοδοι οδηγούν συχνά σε βελτιωμένη ακρίβεια σε σχέση με ένα μόνο δέντρο απόφασης, με τίμημα την αυξημένη πολυπλοκότητα (Coadou, 2016). Τα Ενισχυμένα Δέντρα Αποφάσεων έχουν αποδείξει την ανωτερότητά τους σε ποικίλες εφαρμογές, όπως στην τμηματοποίηση κειμένου, όπου ξεπέρασαν άλλους ταξινομητές σε μη ισορροπημένα σύνολα δεδομένων (Peng et al., 2012), καθώς και στην αξιολόγηση πιστοληπτικής ικανότητας, όπου πέτυχαν μεγαλύτερη ακρίβεια από τις μηχανές υποστήριξης διανυσμάτων (SVMs) και τα νευρωνικά δίκτυα (Bastos, 2022). Πρόσφατες καινοτομίες, όπως τα δέντρα με διανυσματικές τιμές για πολυκλασικές εργασίες και το boosting ανά επίπεδο (layer-by-layer), έχουν βελτιώσει περαιτέρω τη συμπίκνωση των μοντέλων και την ταχύτητα εκπαίδευσης (Ponomareva et al., 2017).

Τυχαία Δάση (Random Forest)

Ο αλγόριθμος Random Forest παρουσιάζεται ως μια ισχυρή μέθοδος ταξινόμησης συλλογικής μάθησης (ensemble) η οποία κατασκευάζει ένα πλήθος δέντρων αποφάσεων κατά τη διάρκεια της εκπαίδευσης και δίνει ως τελική πρόβλεψη την κατηγορία που επιλέγεται από την πλειοψηφία των δέντρων. Ο αλγόριθμος εισάγει τυχαιότητα τόσο επιλέγοντας τυχαία υποσύνολα των δεδομένων για κάθε δέντρο, όσο και τυχαία υποσύνολα χαρακτηριστικών σε κάθε κόμβο διαχωρισμού. Αυτή η διαδικασία μειώνει τη συσχέτιση μεταξύ των δέντρων και ενισχύει τη γενίκευση και τη σταθερότητα του τελικού μοντέλου. Στην παρούσα εργασία επισημαίνεται ότι το Random Forest επιτυγχάνει υψηλή ακρίβεια, ιδιαίτερα σε περιπτώσεις με μεγάλα σύνολα δεδομένων και πολλαπλές μεταβλητές εισόδου, χάρη στην ικανότητά του να διαχειρίζεται θόρυβο και να αποφεύγει την υπερπροσαρμογή. Σε σύγκριση με άλλους αλγόριθμους που εξετάστηκαν, το Random Forest έδειξε σταθερά καλή απόδοση σε διαφορετικές μετρικές και σύνολα δεδομένων, επιβεβαιώνοντας την αξιοπιστία και την ευελιξία του στις εργασίες ταξινόμησης (Leo Breiman, 2001).

Νευρωνικά Δίκτυα

Τα Νευρωνικά δίκτυα είναι υπολογιστικά συστήματα εμπνευσμένα από την δομή και την λειτουργικότητα του ανθρώπινου εγκεφάλου. Αποτελούνται από πολλαπλούς υπολογιστικούς

κόμβους (νευρώνες) οι οποίοι είναι διασυνδεδεμένοι μεταξύ τους όπου επεξεργάζονται τις πληροφορίες συλλογικά και παράλληλα, προσομοιώνοντας τα βιολογικά νευρικά κυκλώματα σε ανθρώπους και ζώα (Kruse et al., 2016). Κάθε τεχνητός νευρώνας δέχεται σήματα εισόδου, τα μετασχηματίζει μέσω μιας συνάρτησης ενεργοποίησης και μεταδίδει το αποτέλεσμα σε άλλους νευρώνες μέσω βαρών σύνδεσης (Awange et al., 2019). Μέσω μιας διαδικασίας που ονομάζεται εκπαίδευση, το δίκτυο προσαρμόζει αυτά τα βάρη με βάση την έκθεση σε δεδομένα, γεγονός που του επιτρέπει να μαθαίνει πολύπλοκα μοτίβα χωρίς να απαιτείται ρητός προγραμματισμός (John G. Taylor, 2002). Η εκπαίδευση των νευρωνικών δικτύων πραγματοποιείται με αλγορίθμους όπως η οπισθοδιάδοση (backpropagation) και η κατάβαση κλίσης (gradient descent), που επιτρέπουν τη ρύθμιση των εσωτερικών βαρών με βάση τα σφάλματα μεταξύ των προβλεπόμενων και των πραγματικών εξόδων (Lote et al., 2020). Ανάλογα με το μαθησιακό υπόδειγμα, τα νευρωνικά δίκτυα μπορούν να λειτουργήσουν:

- Στην επιβλεπόμενη μάθηση, όπου εκπαιδεύονται με επισημασμένα δεδομένα για εργασίες όπως ταξινόμηση και πρόβλεψη. Ένα χαρακτηριστικό παράδειγμα είναι οι πολυεπίπεδοι νευρώνες που χρησιμοποιούν τον αλγόριθμο οπισθοδιάδοσης του λάθους (error backpropagation algorithm) (Lee, Booth & Alam, 2005).
- Στη μη επιβλεπόμενη μάθηση, όπου δεν απαιτούνται επισημασμένα δεδομένα και το δίκτυο ανακαλύπτει κρυμμένες δομές, χρήσιμες σε εφαρμογές όπως η ομαδοποίηση (clustering) και η μείωση διαστάσεων (dimensionality reduction) (Becker, 1991).
- Σε υβριδικά μοντέλα, όπου συνδυάζονται τα δύο υποδείγματα: πρώτα μη επιβλεπόμενη μάθηση για την ανακάλυψη δομών και στη συνέχεια επιβλεπόμενη μάθηση για τη λεπτομερή ρύθμιση της απόδοσης (Hsieh & Chen, 1993).

Χάρη σε αυτές τις ιδιότητες, τα νευρωνικά δίκτυα είναι ιδιαίτερα αποτελεσματικά σε τομείς όπως η ιατρική, η μετεωρολογία, η επεξεργασία εικόνας και η αναγνώριση ομιλίας, ενώ βρίσκουν εφαρμογή σε ένα ευρύ φάσμα προκλήσεων όπου απαιτείται κατανόηση σύνθετων δεδομένων και παροχή βέλτιστων λύσεων (Srivatsa N Joshi et al., 2023).

***k*-Πλησιέστεροι Γείτονες (*k*-Nearest Neighbours (*k*NN))**

Ο αλγόριθμος *k*-Nearest Neighbours (*k*NN) αποτελεί μία απλή αλλά αποτελεσματική μέθοδο μηχανικής μάθησης, η οποία εφαρμόζεται τόσο σε προβλήματα ταξινόμησης όσο και παλινδρόμησης, με κυρίαρχη χρήση στην ταξινόμηση. Ανήκει στις επιβλεπόμενες μεθόδους μάθησης και βασίζεται στην αρχή της ομοιότητας μεταξύ παραδειγμάτων. Η βασική ιδιαιτερότητα είναι ότι δεν προβαίνει σε διαδικασία εκμάθησης κατά τη φάση της εκπαίδευσης αλλά αποθηκεύει τα δεδομένα, αποδίδοντας του χαρακτηριστικά "τεμπέλικου" αλγορίθμου (lazy learner). Όταν εισάγεται ένα νέο, αταξινομητο δείγμα, ο *k*NN υπολογίζει την απόσταση από τα κοντινότερα σημεία του εκπαιδευτικού συνόλου και αναθέτει την ετικέτα της πλειοψηφίας. Παρά την απλότητά του, η επιλογή του κατάλληλου αριθμού γειτόνων *k* επηρεάζει καθοριστικά την ακρίβεια των προβλέψεων. Επιπλέον, η υπολογιστική του επιβάρυνση εντοπίζεται κυρίως στο στάδιο της ταξινόμησης, καθώς κάθε νέο δείγμα

απαιτεί συγκρίσεις με το σύνολο των αποθηκευμένων δεδομένων (Kaushvi Taunk et.al, 2019).

1.3.2 Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Κατηγορίες Προβλημάτων

Οι βασικές κατηγορίες προβλημάτων μη επιβλεπόμενης μάθησης είναι οι παρακάτω:

- **Εξόρυξη κανόνων συσχέτισης:** Η εξόρυξη κανόνων συσχέτισης (Association Rule Mining) είναι μια θεμελιώδης τεχνική της μη επιβλεπόμενης μάθησης που στοχεύει στην ανακάλυψη ουσιαστικών σχέσεων, μοτίβων ή εξαρτήσεων μεταξύ μεταβλητών σε μεγάλα σύνολα δεδομένων. Χρησιμοποιείται ευρέως στην ανάλυση αγοράς (Market-Based-Analysis) με στόχο την ταυτοποίηση συνδυασμών αντικειμένων που εμφανίζονται συχνά στις συναλλαγές του πελάτη π.χ., η ανακάλυψη ότι ένας πελάτης ο οποίος αγοράζει ψωμί, είναι επίσης πολύ πιθανό να αγοράσει και βούτυρο (KJ Cios et.al, 2007). Αυτές οι συσχετίσεις εκφράζονται συνήθως με τη μορφή κανόνων της μορφής $A \rightarrow B$, όπου η παρουσία του στοιχείου A υποδηλώνει αυξημένη πιθανότητα εμφάνισης του στοιχείου B. Η ισχύς και η σημασία αυτών των κανόνων αξιολογούνται με μετρικές όπως η στήριξη (support), η οποία υποδεικνύει πόσο συχνά εμφανίζεται ο κανόνας στο σύνολο δεδομένων η εμπιστοσύνη (confidence), που μετρά την πιθανότητα εμφάνισης του B δεδομένης της παρουσίας του A και ο ενισχυτικός λόγος (lift), ο οποίος συγκρίνει τη συχνότητα συν-εμφάνισης των A και B με εκείνη που θα αναμενόταν αν ήταν στατιστικά ανεξάρτητα (Herrera et al., 2016). Η εξόρυξη κανόνων συσχέτισης έχει αποδειχθεί πολύτιμη και σε άλλους τομείς πέρα από το λιανικό εμπόριο, όπως συστήματα προτάσεων ηλεκτρονικών σεμιναρίων, βιοπληροφορικής και ανίχνευση απάτης πιστωτικών καρτών (Aher & Lobo, 2012) (Sánchez et al., 2009).
- **Ομαδοποίηση:** Η ομαδοποίηση (Clustering) αποτελεί εξίσου μια θεμελιώδη τεχνική της μη επιβλεπόμενης μάθησης και χρησιμοποιείται για να ομαδοποιήσει μη ετικετοποιημένα δεδομένα σε μία ουσιαστική δομή η ομάδα με βάση την ομοιότητά τους. Η επιτυχία της ομαδοποίησης εξαρτάται από τον βαθμό ομοιότητας ή της ανομοιότητας των δεδομένων που χρησιμοποιούνται καθώς αυτός ο βαθμός επηρεάζει τον τρόπο με τον οποίο θα ομαδοποιηθούν τα δεδομένα (Angela Serra & Roberto Tagliaferri, 2018). Επίσης η ομαδοποίηση παίζει σημαντικό ρόλο σε διάφορους τομείς μελέτης όπως η ιατρική διάγνωση, η ανάκτηση πληροφοριών, το μάρκετινγκ και οι

κοινωνικές επιστήμες. Για παράδειγμα, στην ιατρική διάγνωση μπορεί να βοηθήσει στην κατανομή των ασθενών σε διαφορετικές ομάδες βάσει την κατάσταση της ασθένειάς τους ενώ στο μάρκετινγκ μπορεί να συμβάλλει στον εντοπισμό πελατών με παρόμοια ενδιαφέροντα με σκοπό την αύξηση αποτελεσματικότητας των διαφημίσεων (Meshal Shutaywi, Nezamoddin N. Kachouie, 2021)

Κοινοί Αλγόριθμοι

Self-Organizing Map

Ο αλγόριθμος Self-Organizing Map (SOM) είναι μια τεχνική μη επιβλεπόμενης μάθησης που χαρτογραφεί δεδομένα υψηλής διάστασης σε ένα χαμηλότερης διάστασης πλέγμα, συνήθως δισδιάστατο, διατηρώντας τις τοπολογικές σχέσεις του αρχικού χώρου εισόδου. Αυτό καθιστά τους SOM ιδιαίτερα αποτελεσματικά εργαλεία για ομαδοποίηση και απεικόνιση σύνθετων συνόλων δεδομένων χωρίς την ανάγκη για προϋπάρχουσες ετικέτες. Κάθε μονάδα ή νευρώνας στον χάρτη ανταγωνίζεται για να εκπροσωπήσει καλύτερα ένα δεδομένο πρότυπο εισόδου, και οι γειτονικοί νευρώνες ενημερώνονται ταυτόχρονα, επιτρέποντας τον σχηματισμό οργανωμένων και νοηματικών ομάδων. Οι SOM έχουν ευρεία εφαρμογή στην ανάλυση εγγράφων, μικρο-ανάλυση δεδομένων και ανάλυση περιήγησης στον ιστό χάρη στην ικανότητά τους να αποκαλύπτουν πρότυπα και να μειώνουν τη διάσταση των δεδομένων (Cialfi, 2019). Η ευελιξία του αλγορίθμου έχει οδηγήσει σε επεκτάσεις όπως οι τυχαιοποιημένοι αυτοοργανούμενοι χάρτες (Randomized SOMs), οι οποίοι βελτιώνουν την προσαρμοστικότητα και την απόδοση κατά την ανάλυση δεδομένων υψηλής διάστασης και τοπολογικά πολύπλοκων δομών (Rougier & Detorakis, 2020).

Εββιανή Μάθηση

Η Εββιανή μάθηση αποτελεί ένα θεμελιώδες πρότυπο μη επιβλεπόμενης μάθησης στη νευροεπιστήμη και στα τεχνητά νευρωνικά δίκτυα, βασιζόμενο στην αρχή ότι η σύνδεση μεταξύ δύο νευρώνων ενισχύεται όταν αυτοί ενεργοποιούνται ταυτόχρονα. Ο κανόνας αυτός, που συνοψίζεται συχνά στη φράση «νευρώνες που ενεργοποιούνται μαζί, συνδέονται μαζί», αποτυπώνει τη συσχέτιση μεταξύ της προσυναπτικής και της μετασυναπτικής δραστηριότητας. Σε πρακτική εφαρμογή, η Εββιανή μάθηση ενημερώνει το συναπτικό βάρος μεταξύ δύο νευρώνων ανάλογα με το γινόμενο των ενεργοποιήσεών τους, γεγονός που οδηγεί φυσικά στην ενίσχυση των προτύπων και των συσχετίσεων που υπάρχουν στα δεδομένα εισόδου (Chakraverty et al., 2019). Παραλλαγές της Εββιανής μάθησης, όπως η ανταγωνιστική Εββιανή μάθηση, προσαρμόζουν τις μεταβολές των βαρών λαμβάνοντας υπόψη τη δραστηριότητα των ανταγωνιζόμενων κόμβων, επιτρέποντας στα νευρωνικά δίκτυα να μαθαίνουν διακριτά χαρακτηριστικά ή πρότυπα εντός ενός συνόλου δεδομένων (White, 1992).

Αλγόριθμος Apriori

Ο αλγόριθμος Apriori είναι ένας κλασικός αλγόριθμος που χρησιμοποιείται στην εξόρυξη κανόνων συσχέτισης, η βασική του ιδέα περιλαμβάνει την αξιοποίηση προηγούμενων

γνώσεων για την επαναληπτική δημιουργία υποψήφιων συνόλων αντικειμένων και τον αποκλεισμό των μη συχνών συνόλων (Shen et al, 2024). Ο αλγόριθμος Apriori λειτουργεί με μια δομημένη, επαναληπτική διαδικασία για την εξαγωγή συχνών συνόλων αντικειμένων από βάσεις δεδομένων συναλλαγών. Αρχικά, δημιουργεί υποψήφια σύνολα αντικειμένων μέσω συνδυασμών των συχνών ($k-1$) συνόλων που εντοπίστηκαν στον προηγούμενο κύκλο, εξαίρωντας όμως κάθε υποψήφιο του οποίου οποιοδήποτε υποσύνολο μήκους $(k-1)$ δεν είναι ήδη συχνό. Αυτή η τεχνική, γνωστή ως «κλάδεμα υποψηφίων», συμβάλλει στην αποδοτικότητα του αλγορίθμου, μειώνοντας περιττούς υπολογισμούς. Στη συνέχεια, υπολογίζεται η υποστήριξη (support) κάθε υποψηφίου συνόλου μέσω πλήρους σάρωσης της βάσης δεδομένων, προκειμένου να μετρηθεί ο αριθμός συναλλαγών που περιέχουν όλα τα αντικείμενα του υποψηφίου συνόλου. Αν και απαιτητική υπολογιστικά, αυτή η διαδικασία είναι απαραίτητη. Έπειτα, εντοπίζονται τα υψηλής συχνότητας πρότυπα, δηλαδή εκείνα τα σύνολα των οποίων η υποστήριξη υπερβαίνει ένα προκαθορισμένο όριο. Εφόσον τέτοια πρότυπα υπάρχουν, διατηρούνται ως συχνά σύνολα και η διαδικασία συνεχίζεται με αύξηση του μεγέθους των συνόλων κατά μία μονάδα ($k+1$). Αν δεν εντοπιστούν νέα συχνά σύνολα, ο αλγόριθμος τερματίζεται. Αυτή η βήμα προς βήμα μεθοδολογία αναδεικνύει τις βασικές αρχές του Apriori: επαναληπτική επέκταση, φιλτράρισμα βάσει υποστήριξης και διακοπή όταν δεν εντοπίζονται νέες συχνότητες (Suprianto Panjaitan et al 2019).

***k*-means**

Ο αλγόριθμος *k*-means χρησιμοποιείται για τη διαμέριση συνόλων δεδομένων σε *k* διακριτές ομάδες (clusters), όπου κάθε σημείο δεδομένων ανήκει στο σύμπλεγμα του οποίου το μέσο είναι το πλησιέστερο. Η απλότητα και η αποδοτικότητά του τον έχουν καταστήσει ως έναν από τους πλέον διαδεδομένους αλγορίθμους ομαδοποίησης στον τομέα της βιοπληροφορικής, μάρκετινγκ, εγκληματολογική ανάλυση κ.α (Bao, 2021). Ωστόσο, παρά τη δημοτικότητά του, ο αλγόριθμος παρουσιάζει ορισμένους περιορισμούς, όπως η τυχαία αρχικοποίηση των κεντροειδών, η οποία μπορεί να οδηγήσει σε απρόβλεπτη σύγκλιση. Επιπλέον, απαιτείται ο προκαθορισμός του αριθμού των συστάδων, γεγονός που επηρεάζει τη μορφή των παραγόμενων συστάδων και την ευαισθησία σε ακραίες τιμές (Mohiuddin Ahmed et.al 2020). Οι βελτιώσεις στον αλγόριθμο *k*-means έχουν επικεντρωθεί στη βελτιστοποίηση της αρχικοποίησης των κεντροειδών, στη βελτίωση της διαχείρισης του θορύβου, καθώς και στη μείωση της υπολογιστικής πολυπλοκότητας για μεγάλα σύνολα δεδομένων (Leela, 2011).

Ιεραρχική Ομαδοποίηση (Hierarchical clustering)

Η βασική ιδέα των αλγορίθμων ιεραρχικής ομαδοποίησης είναι η κατασκευή ιεραρχικών σχέσεων μεταξύ των δεδομένων με βάση κάποιο μέτρο ομοιότητας ή ανομοιότητας μεταξύ των ομάδων. Η ιεραρχική μέθοδος ομαδοποίησης διακρίνεται σε δύο τύπους, τη διαιρετική ομαδοποίηση (Divisive Clustering) και τη συσσωρευτική ομαδοποίηση (Agglomerative Clustering) (Xingcheng Ran et.al 2022). Η διαιρετική ιεραρχική ομαδοποίηση (Divisive Hierarchical Clustering) είναι ευρέως γνωστή ως ομαδοποίηση από πάνω προς τα κάτω, καθώς ξεκινά με όλα τα αντικείμενα συγκεντρωμένα σε ένα ενιαίο σύνολο. Μέσω διαδοχικών επαναλήψεων, το αρχικό σύνολο διαχωρίζεται προοδευτικά σε μικρότερα υποσύνολα, συχνά με τη χρήση του αλγορίθμου k -means. Η διαδικασία συνεχίζεται μέχρι κάθε αντικείμενο να ανήκει σε ξεχωριστό σύνολο ή μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού. Η μέθοδος αυτή είναι άκαμπτη, δηλαδή κάθε συγχώνευση ή διαχωρισμός που πραγματοποιείται δεν μπορεί να αναιρεθεί. Για να είναι τα αποτελέσματα της ομαδοποίησης χρήσιμα, πρέπει να είναι ερμηνεύσιμα, κατανοητά και πρακτικά αξιοποιήσιμα. Επιπλέον, είναι σημαντικό να απαιτούνται ελάχιστες γνώσεις του πεδίου εφαρμογής για τον καθορισμό των απαραίτητων παραμέτρων εισόδου, ενώ η μέθοδος χαρακτηρίζεται από ανθεκτικότητα ως προς τη σειρά εισαγωγής των δεδομένων (M. Venkat Reddy et.al 2017). Η συσσωρευτική ιεραρχική ομαδοποίηση είναι μια διαδικασία ομαδοποίησης που ακολουθεί προσέγγιση από κάτω προς τα πάνω. Αρχικά, κάθε δεδομένο θεωρείται ως μια ξεχωριστή συστάδα. Σε κάθε επόμενο βήμα, εντοπίζονται οι δύο συστάδες που είναι πιο κοντά μεταξύ τους και συγχωνεύονται. Η διαδικασία αυτή επαναλαμβάνεται μέχρι όλα τα δεδομένα να ανήκουν τελικά σε μία και μοναδική συστάδα. Η μέθοδος αυτή δημιουργεί μια ιεραρχική δομή συστάδων, στην οποία κάθε επίπεδο αντιπροσωπεύει διαφορετικό βαθμό συγγένειας μεταξύ ομάδων δεδομένων. Τέτοιου τύπου ιεραρχίες είναι ιδιαίτερα χρήσιμες σε περιπτώσεις όπου ενδιαφέρει η μελέτη της σταδιακής συνάφειας μεταξύ των δεδομένων ή όταν ο τελικός αριθμός συστάδων δεν είναι εκ των προτέρων γνωστός (Marcel R. Ackermann et.al 2012).

Πιθανοτική Ομαδοποίηση (Probabilistic Clustering)

Στην πιθανοτική ομαδοποίηση, τα σημεία δεδομένων ομαδοποιούνται με βάση συναρτήσεις πυκνότητας πιθανότητας (Probability Density Functions - PDFs) που αποτυπώνουν την αβεβαιότητα και τη φύση της κατανομής των δεδομένων. Σε αντίθεση με τις παραδοσιακές μεθόδους, όπως το k -means, οι οποίες βασίζονται σε αυστηρές αναθέσεις (hard assignments), οι πιθανοτικές προσεγγίσεις επιτρέπουν πιο "ήπιες" αναθέσεις (soft assignments), αντανακλώντας την πιθανότητα συμμετοχής σε κάθε ομάδα (cluster). Ένα αποτελεσματικό μοντέλο χρησιμοποιεί την εκτίμηση πυκνότητας πιθανότητας βασισμένη σε wavelet για τον εντοπισμό προτύπων ομαδοποίησης, ειδικά σε σύνθετα και θορυβώδη δεδομένα, όπως τα προφίλ γονιδιακής έκφρασης κατά την εξέλιξη του καρκίνου (Kordestani et al., 2016). Μια άλλη μελέτη εισήγαγε μια αναδρομική τεχνική επικύρωσης χρησιμοποιώντας αναπαραστάσεις χώρου-κλίμακας της συνάρτησης πυκνότητας πιθανότητας των δεδομένων, η οποία επιτρέπει τη στατιστικά ισχυρή ανίχνευση συστάδων σε πολλαπλά επίπεδα ανάλυσης (Sakai et al., 2007). Επιπρόσθετα, έχουν προταθεί τεχνικές βασισμένες σε εκτιμήσεις τρόπων της συνάρτησης πυκνότητας πιθανότητας μέσω bootstrapping, για τον αξιόπιστο προσδιορισμό του αριθμού των συστάδων χωρίς την παραδοχή κανονικότητας των

δεδομένων (Herbin & Bonnet, 2002). Αυτές οι προσεγγίσεις καταδεικνύουν πώς τα πιθανοτικά πλαίσια μπορούν να διαχειριστούν την ασάφεια των δεδομένων και τη δομική πολυπλοκότητα πιο αποτελεσματικά από τις ντετερμινιστικές μεθόδους.

1.3.3 Ενισχυτική Μάθηση (Reinforcement Learning)

Οι βασικές κατηγορίες προβλημάτων της ενισχυτικής μάθησης είναι οι παρακάτω:

- **Εξερεύνηση vs. Αξιοποίηση (Exploration vs. Exploitation):** Το δίλημμα ανάμεσα στο να δοκιμάζει ο πράκτορας νέες ενέργειες για ενδεχόμενα υψηλότερες ανταμοιβές (εξερεύνηση) και στο να επιλέγει ενέργειες που ήδη ξέρει ότι λειτουργούν καλά (αξιοποίηση) είναι βασικό σε όλα τα προβλήματα ενισχυτικής μάθησης (Kaelbling et al., 1995).
- **Μερική Παρατηρησιμότητα (Partial Observability):** Ο πράκτορας δεν έχει πλήρη εικόνα της κατάστασης του περιβάλλοντος, κάτι που απαιτεί τεχνικές όπως POMDPs (Partially Observable Markov Decision Processes) (Kaelbling et al., 1995)
- **Εκμάθηση από Καθυστερημένη Ανταμοιβή (Delayed Rewards):** Πολλές ενέργειες έχουν αποτέλεσμα μακροπρόθεσμα, γεγονός που δυσκολεύει την απόδοση σωστής αξίας στις πράξεις του πράκτορα (Sutton, 1992)
- **Γενίκευση & Ιεραρχία (Generalization & Hierarchical – RL):** Ο πράκτορας πρέπει να μπορεί να γενικεύσει από παρόμοιες εμπειρίες και να χρησιμοποιεί ιεραρχίες ενεργειών για να χειρίζεται μεγάλα χρονικά εύρη προβλημάτων (Kaelbling et al., 1995), (Szepesvári, 2010).
- **Αποδοτικότητα Δεδομένων (Sample Efficiency):** Σε πολλά πραγματικά προβλήματα, ο πράκτορας έχει περιορισμένες ευκαιρίες για αλληλεπίδραση με το περιβάλλον, άρα απαιτούνται αλγόριθμοι με υψηλή αποδοτικότητα δεδομένων (Szepesvári, 2009)
- **Αντιμετώπιση Θορυβώδους ή Αβέβαιου Περιβάλλοντος (Stochastic Environments):** Τα περιβάλλοντα συχνά έχουν στοχαστική φύση, δηλαδή ίδιες ενέργειες μπορεί να οδηγούν σε διαφορετικά αποτελέσματα (Diederichs, 2019)
- **Επαναλαμβανόμενα vs. Επεισοδικά Προβλήματα (Continuing vs. Episodic Tasks):** Τα προβλήματα μπορεί να είναι συνεχόμενα χωρίς σαφές τέλος ή να χωρίζονται σε επεισόδια με αρχή και τέλος (Sutton & Barto, 1998)

- **Μοντελοβασισζόμενη vs. Μη Μοντελοβασισζόμενη Μάθηση (Model-based vs. Model-free RL):** Η διαφορά μεταξύ των μεθόδων που δημιουργούν εσωτερικό μοντέλο του περιβάλλοντος και αυτών που μαθαίνουν απευθείας από εμπειρία (Ghasemi & Ebrahimi, 2024).

Κοινοί Αλγόριθμοι

Q-Learning

Ο Q-learning είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος ενισχυτικής μάθησης χρονικής διαφοράς που επιτρέπει στους πράκτορες να μαθαίνουν βέλτιστες συμπεριφορές μεγιστοποιώντας τις σωρευτικές ανταμοιβές. Ενώ το παραδοσιακό Q-learning, συχνά χρησιμοποιώντας πίνακες αναζήτησης, έχει αποδειχθεί ότι συγκλίνει στη βέλτιστη λύση, μπορεί να παρουσιάσει αστάθεια όταν χρησιμοποιεί προσέγγιση συνάρτησης τιμής, όπως βαθιά νευρωνικά δίκτυα, ή όταν λειτουργεί σε στοχαστικά περιβάλλοντα. Για να αντιμετωπιστούν αυτές οι προκλήσεις και να βελτιωθεί η απόδοση, έχουν αναπτυχθεί βελτιώσεις όπως το Multi Q-learning (Duryea et al., 2016). Παραλλαγές, όπως το Q-learning που βασίζεται σε σχετική ανταμοιβή (Relative Reward-Based Q-learning), στοχεύουν στη μείωση του χρόνου εκμάθησης, δίνοντας έμφαση στις συγκρίσεις άμεσης ανταμοιβής μεταξύ των καταστάσεων (Pandey et al., 2010).

Μάθηση Χρονικής Διαφοράς (Temporal Difference Learning)

Η Μάθηση Χρονικής διαφοράς (Temporal Difference - TD) αποτελεί μια θεμελιώδη προσέγγιση στην ενισχυτική μάθηση. Συνδυάζει στοιχεία από τις μεθόδους Monte Carlo και τον δυναμικό προγραμματισμό για την εκτίμηση συναρτήσεων αξίας. Αυτό επιτυγχάνεται μέσω της άμεσης εκμάθησης από ανεπεξέργαστη εμπειρία. Οι μέθοδοι χρονικής διαφοράς ανανεώνουν τις προβλέψεις τους βασισζόμενες στη διαφορά μεταξύ διαδοχικών εκτιμήσεων, καθιστώντας τις ιδανικές για συνεχή, επαυξητική μάθηση. Μια ιδιαίτερα ισχυρή παραλλαγή της, η TD(λ), ενσωματώνει ίχνη επιλεξιμότητας (eligibility traces). Αυτά τα ίχνη συμβάλλουν στην κατανομή της "πίστωσης" σε προηγούμενες καταστάσεις, βασισζόμενα σε έναν παράγοντα απόσβεσης λ , γεγονός που βελτιώνει την αποτελεσματικότητα της μάθησης (Kunz 2013). Οι μέθοδοι χρονικής διαφοράς αποτελούν τη βάση για σημαντικούς αλγορίθμους όπως ο Q-Learning και οι μέθοδοι Actor-Critic (Cichosz, 1994). Επιπλέον, έχουν επεκταθεί για να αντιμετωπίσουν προκλήσεις όπως ο θόρυβος διάδοσης τιμής (Value Propagation Noise) και η ανισορροπία στην επίσκεψη καταστάσεων (State Visitation Imbalance), μέσω νεότερων παραλλαγών όπως η διακριτική μάθηση TD (DTD) (Ma, 2023). Η βιολογική σημασία της μάθησης χρονικής διαφοράς τεκμηριώνεται επίσης από την αντιστοιχία της με τα ντοπαμινεργικά σήματα πρόβλεψης ανταμοιβής που παρατηρούνται στον εγκέφαλο. (Kurth-Nelson & Redish, 2009).

Γενετικοί Αλγόριθμοι

Ο Γενετικοί αλγόριθμοι αποτελούν μια μέθοδο βελτιστοποίησης, κατάλληλη για την επίλυση πολύπλοκων προβλημάτων, βασισμένη στην αρχή της γενετικής επιλογής. Πέρα από τις εφαρμογές τους στην βελτιστοποίηση, εξυπηρετούν επίσης τους σκοπούς της μηχανικής μάθησης, καθώς και την έρευνα και ανάπτυξη. Λειτουργούν αναλογικά με τις βιολογικές διεργασίες δημιουργίας χρωμοσωμάτων, όπου ενέργειες όπως η επιλογή, η διασταύρωση και η μετάλλαξη συνιστούν τις γενετικές λειτουργίες όπου εφαρμόζονται αρχικά σε έναν τυχαίο πληθυσμό (Lambora et.al 2019). Στους γενετικούς αλγόριθμους, οι δυαδικές αλφαριθμητικές ακολουθίες αποθηκεύονται στη μνήμη ενός υπολογιστή και, με την πάροδο του χρόνου, τροποποιούνται με τρόπο ανάλογο με τον οποίο εξελίσσονται οι πληθυσμοί ατόμων μέσω της φυσικής επιλογής. Παρόλο που το υπολογιστικό περιβάλλον είναι ιδιαίτερα απλουστευμένο σε σύγκριση με τον φυσικό κόσμο, οι γενετικοί αλγόριθμοι είναι ικανοί να εξελίξουν εκπληκτικά πολύπλοκες και ενδιαφέρουσες δομές. Αυτές οι δομές, που αποκαλούνται "άτομα" ή "λύσεις", μπορούν να αναπαραστήσουν λύσεις σε προβλήματα, στρατηγικές για παιχνίδια, οπτικές εικόνες ή ακόμη και προγράμματα υπολογιστών (Forrest, 1996).

Monte Carlo

Οι μέθοδοι Monte Carlo αποτελούν μια κατηγορία υπολογιστικών αλγορίθμων που αξιοποιούν την επανειλημμένη τυχαία δειγματοληψία για την παραγωγή αριθμητικών αποτελεσμάτων. Χρησιμοποιούνται συχνά για την επίλυση προβλημάτων τα οποία, αν και θεωρητικά ντετερμινιστικά, είναι υπερβολικά περίπλοκα για αναλυτικές λύσεις. Οι τεχνικές αυτές έχουν τις ρίζες τους στη θεωρία πιθανοτήτων και εφαρμόζονται ευρέως σε εργασίες όπως η εκτίμηση ολοκληρωμάτων και η προσομοίωση φυσικών συστημάτων. Η κεντρική ιδέα συνίσταται στη μοντελοποίηση ενός προβλήματος ως μια πιθανολογική διαδικασία και στη χρήση τυχαίων δειγμάτων για την προσέγγιση της λύσης του. Για παράδειγμα, η εκτίμηση ενός ολοκληρώματος μπορεί να αναδιατυπωθεί ως η εκτίμηση μιας αναμενόμενης τιμής, αντλώντας τυχαία δείγματα από μια κατάλληλη κατανομή και υπολογίζοντας τον μέσο όρο των αποτελεσμάτων (Cragg, 1990)(Mañá, 2017). Αυτές οι μέθοδοι είναι ευέλικτες, επιτρέποντας την εφαρμογή τους ακόμα και σε συστήματα που δεν διαθέτουν ενσωματωμένο τυχαίο χαρακτήρα, αρκεί να μπορεί να δοθεί μια πιθανολογική ερμηνεία. Ένα κλασικό παράδειγμα από τον 18ο αιώνα, περιλαμβάνει την τυχαία ρίψη μιας βελόνας για την εκτίμηση της τιμής του π . Αυτό καταδεικνύει σαφώς πώς οι στοχαστικές προσομοιώσεις μπορούν να προσφέρουν σημαντικές γνώσεις σε καθαρά μαθηματικά προβλήματα(Mascagni & Simonov, 2004).

Proximal Policy Optimization (PPO)

Ο αλγόριθμος Proximal Policy Optimization αποτελεί μια από τις πιο διαδεδομένες μεθόδους ενισχυτικής μάθησης, σχεδιασμένος ώστε να εξισορροπεί την αποδοτικότητα των ενημερώσεων της πολιτικής με τη σταθερότητα και την αξιοπιστία της εκπαίδευσης. Εισήχθη ως μια απλούστερη, εναλλακτική λύση σε σχέση με τον Trust Region Policy Optimization (TRPO) και χρησιμοποιεί μια τροποποιημένη αντικειμενική συνάρτηση για να περιορίσει τις αλλαγές στην πολιτική, αποτρέποντας έτσι απότομες μεταβολές και εξασφαλίζοντας πιο σταθερή μαθησιακή διαδικασία (Schulman et al., 2017). Ο PPO έχει επιδείξει εξαιρετικά

αποτελέσματα σε ποικιλία πολύπλοκων προβλημάτων ελέγχου, όπως η ρομποτική κίνηση και τα παιχνίδια Atari.

Παρά τα πλεονεκτήματά του, εξακολουθούν να υπάρχουν προκλήσεις, όπως η σχετικά χαμηλή αποδοτικότητα στη χρήση δειγμάτων (sample inefficiency) και η πιθανότητα υπερπροσαρμογής (overfitting) κατά τη διάρκεια παρατεταμένης εκπαίδευσης (Yu et al., 2021). Πρόσφατες βελτιώσεις, όπως ο Truly PPO και ο PPO με μηχανισμούς ανατροφοδότησης πολιτικής (policy feedback), στοχεύουν στη διόρθωση περιορισμών του αρχικού μηχανισμού clipping και στην περαιτέρω ενίσχυση της σταθερότητας της πολιτικής (Wang et al., 2019; Gu et al., 2022). Παράλληλα, παραλλαγές όπως τα P3O και APPO επεκτείνουν τον PPO προς την κατεύθυνση της «ασφαλούς» ενισχυτικής μάθησης (safe reinforcement learning), ενσωματώνοντας μηχανισμούς διαχείρισης περιορισμών για να διασφαλίζεται η ασφάλεια και η κανονιστική συμμόρφωση σε εφαρμογές υψηλού ρίσκου (Zhang et al., 2022; Dai et al., 2023).

Συνοψίζοντας, οι βασικές προσεγγίσεις της μηχανικής μάθησης διακρίνονται με βάση το επίπεδο εποπτείας στη φάση της εκπαίδευσης, με την ενισχυτική μάθηση να διαφοροποιείται εννοιολογικά και μεθοδολογικά από τις επιβλεπόμενες και μη επιβλεπόμενες τεχνικές. Η φύση της ενισχυτικής μάθησης η οποία βασίζεται στη λήψη αποφάσεων μέσω αλληλεπίδρασης πράκτορα-περιβάλλοντος και στη σταδιακή ενίσχυση μέσω ανταμοιβών την καθιστά ιδιαίτερα κατάλληλη για προβλήματα που εντοπίζονται σε δυναμικά, μη στατικά περιβάλλοντα, όπως αυτά που συναντώνται σε εφαρμογές παιχνιδιών. Στο επόμενο κεφάλαιο εξετάζονται σε βάθος οι κυριότεροι αλγόριθμοι ενισχυτικής μάθησης, με στόχο την αποσαφήνιση των τεχνικών αρχών που διέπουν τη λειτουργία τους, των διαφορών μεταξύ τους, καθώς και των πλεονεκτημάτων και περιορισμών τους, προετοιμάζοντας το έδαφος για την τεκμηριωμένη επιλογή της κατάλληλης μεθοδολογίας στο πλαίσιο της παρούσας εφαρμογής.

ΚΕΦΑΛΑΙΟ 2: ΑΛΓΟΡΙΘΜΟΙ ΚΑΙ ΘΕΜΕΛΙΩΔΕΙΣ ΕΝΝΟΙΕΣ ΕΝΙΣΧΥΤΙΚΗΣ ΜΑΘΗΣΗΣ

Οι αλγόριθμοι Ενισχυτικής Μάθησης σχεδιάζονται για να βοηθήσουν τους πράκτορες να αναπτύξουν βέλτιστες συμπεριφορές μέσω αλληλεπιδράσεων με το περιβάλλον τους, με σκοπό τη μεγιστοποίηση των σωρευτικών ανταμοιβών. Αυτοί οι αλγόριθμοι κατηγοριοποιούνται περαιτέρω σε μεθόδους βασισμένες στην αξία, στην πολιτική, ή Actor-Critic. Μεταξύ των κλασικών αλγορίθμων που βασίζονται στην αξία συγκαταλέγονται ο Q-Learning και ο SARSA, οι οποίοι εκτιμούν την αξία των ενεργειών για την υποβοήθηση της λήψης αποφάσεων (AlMahamid & Grolinger, 2021). Οι μέθοδοι βασισμένες σε πολιτική, μαθαίνουν απευθείας τη συνάρτηση πολιτικής η οποία αντιστοιχίζει τις καταστάσεις σε ενέργειες. Οι μέθοδοι Actor-Critic συνδυάζουν αμφότερες τις στρατηγικές για βελτιωμένη σταθερότητα και αποδοτικότητα. Με την άνοδο της βαθιάς ενισχυτικής μάθησης, η οποία ενσωματώνει νευρωνικά δίκτυα σε αυτά τα μοντέλα, έχουν καταστεί εφικτοί αλγόριθμοι όπως ο Soft Actor-Critic (SAC), οι οποίοι επιδεικνύουν υψηλές επιδόσεις σε απαιτητικά περιβάλλοντα, όπως οι προσομοιώσεις MuJoCo (Pradhan, 2024). Πιο πρόσφατα, έχουν αναπτυχθεί προσεγγίσεις μετα-μάθησης (meta-learning) με στόχο την αυτόματη ανακάλυψη νέων αλγορίθμων ενισχυτικής μάθησης. Χαρακτηριστικό παράδειγμα αποτελεί ο Learned Policy Gradient (LPG), ο οποίος επιδεικνύει αξιοσημείωτη ικανότητα γενίκευσης σε άγνωστες εργασίες (Oh et al., 2020), (Co-Reyes et al., 2021). Οι εξελίξεις αυτές υπογραμμίζουν την ευελιξία της ενισχυτικής μάθησης σε ένα ευρύ φάσμα εφαρμογών, από τον τομέα των ηλεκτρονικών παιχνιδιών έως αυτόν της ρομποτικής.

2.1 Θεμελιώδεις Αρχές Ενισχυτικής Μάθησης

Τα δύο καθοριστικά χαρακτηριστικά της ενισχυτικής μάθησης είναι η μάθηση μέσω δοκιμής και σφάλματος και η καθυστερημένη ανταμοιβή όπου οι ενέργειες δεν επηρεάζουν μόνο τα άμεσα αποτελέσματα, αλλά και τις μελλοντικές καταστάσεις και τα μακροπρόθεσμα κέρδη (Sutton, 1992). Ο πυρήνας της ενισχυτικής μάθησης περιλαμβάνει έναν κυκλικό μηχανισμό όπου ο πράκτορας παρατηρεί μια κατάσταση, επιλέγει μια ενέργεια βασιζόμενος σε μια πολιτική, λαμβάνει μια ανταμοιβή και μεταβαίνει σε μια νέα κατάσταση. Αυτός ο βρόχος συνεχίζεται καθώς ο πράκτορας επιδιώκει να βελτιώσει την πολιτική του μέσω της ανατροφοδότησης (Ghasemi & Ebrahimi, 2024). Τα βασικά συστατικά στοιχεία της ενισχυτικής μάθησης περιλαμβάνουν τον πράκτορα, το περιβάλλον, το σήμα ανταμοιβής, την πολιτική και τις συναρτήσεις αξίας, τα οποία όλα μαζί σχηματίζουν το πρόβλημα μάθησης μαθηματικά ως μια Διαδικασία Μαρκόφ (Markov Decision Process - MDP) (Even-Dar, 2016). Αυτές οι αρχές αποτελούν τη βάση τόσο για βασικές μεθόδους, όπως η Q-Learning, όσο και για προηγμένες τεχνικές βαθιάς ενισχυτικής μάθησης. Αυτές έχουν εφαρμοστεί με επιτυχία σε διάφορους τομείς, όπως η ρομποτική, τα ηλεκτρονικά παιχνίδια και η υγειονομική περίθαλψη (Shimpi, 2025).

2.2 Βασικοί Αλγόριθμοι και Δομή Λειτουργίας

2.2.1 Q-Learning

Ο αλγόριθμος Q-learning είναι μια θεμελιώδης μέθοδος στην ενισχυτική μάθηση, η οποία επιτρέπει στους πράκτορες να μαθαίνουν πώς να δρουν βέλτιστα σε ελεγχόμενα Μαρκοβιανά περιβάλλοντα. Αναπτύχθηκε αρχικά από τον Watkins (1989) και έχει έκτοτε καθιερωθεί ως ένα ισχυρό εργαλείο τόσο στην τεχνητή νοημοσύνη όσο και στη στατιστική, ειδικά στην εξαγωγή βέλτιστων στρατηγικών αποφάσεων.

Στον πυρήνα του, ο Q-learning είναι ένας επαυξητικός αλγόριθμος για την εκτίμηση μιας βέλτιστης στρατηγικής αποφάσεων σε προβλήματα αποφάσεων άπειρου ορίζοντα. Μπορεί να θεωρηθεί ως μια μέθοδος ασύγχρονου δυναμικού προγραμματισμού. Η βασική ιδέα είναι η διαδοχική βελτίωση των εκτιμήσεων της ποιότητας συγκεκριμένων ενεργειών σε συγκεκριμένες καταστάσεις.

Ο Q-learning λειτουργεί σε περιβάλλοντα που μπορούν να μοντελοποιηθούν ως διαδικασίες αποφάσεων Markov (Markov Decision Processes - MDPs). Ένα MDP ορίζεται από ένα σύνολο καταστάσεων S , ένα σύνολο ενεργειών A , μια συνάρτηση μετάβασης $P(s'|s, a)$ που δίνει την πιθανότητα μετάβασης στην κατάσταση s' από την κατάσταση s με την εκτέλεση της ενέργειας a , και μια συνάρτηση ανταμοιβής $R(s, a, s')$ που δίνει την άμεση ανταμοιβή για τη μετάβαση από την κατάσταση s στην s' μέσω της ενέργειας a .

Η κεντρική έννοια στον Q-learning είναι η συνάρτηση τιμής $Q(s, a)$, η οποία αντιπροσωπεύει την αναμενόμενη συνολική μελλοντική ανταμοιβή που μπορεί να επιτευχθεί ξεκινώντας από την κατάσταση s , εκτελώντας την ενέργεια a , και στη συνέχεια ακολουθώντας μια βέλτιστη στρατηγική. Η συνάρτηση τιμής Q ικανοποιεί την εξίσωση Bellman βέλτιστης τιμής:

$$Q^i(s, a) = E[R_{t+1} + \gamma \cdot \max_{a'} Q^i(S_{t+1}, a') \mid S_t = s, A_t = a]$$

όπου $Q^*(s, a)$ είναι η βέλτιστη τιμή Q , R_{t+1} είναι η άμεση ανταμοιβή, $\gamma \in [0, 1)$ είναι ο συντελεστής έκπτωσης, και $\max_{a'} Q^*(S_{t+1}, a')$ είναι η μέγιστη βέλτιστη τιμή Q στην επόμενη κατάσταση S_{t+1} . Ο συντελεστής έκπτωσης καθορίζει τη σημασία των μελλοντικών ανταμοιβών. Μια τιμή γ κοντά στο 0 εστιάζει στις άμεσες ανταμοιβές, ενώ μια τιμή κοντά στο 1 δίνει μεγαλύτερη βαρύτητα στις μακροπρόθεσμες ανταμοιβές. (Watkins & Dayan, 1992; Clifton & Lauer, 2020).

Ο Q-learning είναι ένας αλγόριθμος χωρίς μοντέλο (model-free), πράγμα που σημαίνει ότι δεν απαιτεί γνώση των μεταβατικών συναρτήσεων ή ανταμοιβής του περιβάλλοντος. Μαθαίνει τις βέλτιστες τιμές Q μέσω της εμπειρίας, αλληλεπιδρώντας με το περιβάλλον.

Ο αλγόριθμος ενημέρωσης του Q-learning δίνεται από τον ακόλουθο κανόνα

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t)]$$

όπου:

- $Q(S_t, A_t)$ είναι η τρέχουσα εκτίμηση της τιμής Q για την κατάσταση S_t και την ενέργεια A_t .

- $\alpha \in (0, 1]$ είναι ο ρυθμός μάθησης (learning rate), ο οποίος καθορίζει πόσο γρήγορα ο παράγοντας προσαρμόζεται στις νέες πληροφορίες.
- R_{t+1} είναι η άμεση ανταμοιβή που λαμβάνεται μετά την εκτέλεση της ενέργειας A_t στην κατάσταση S_t και τη μετάβαση στην κατάσταση S_{t+1} .
- γ είναι ο συντελεστής έκπτωσης
- $\max_a Q(S_{t+1}, a')$ είναι η μέγιστη τιμή Q που μπορεί να επιτευχθεί στην επόμενη κατάσταση S_{t+1} με την εκτέλεση οποιασδήποτε ενέργειας a'

Αυτός ο κανόνας ενημέρωσης είναι μια μορφή μάθησης χρονικής διαφοράς (temporal difference learning). Ο όρος $[R_{t+1} + \gamma \max_a Q(S_{t+1}, a') - Q(S_t, A_t)]$ είναι το σφάλμα χρονικής διαφοράς το οποίο αντιπροσωπεύει την διαφορά μεταξύ της τρέχουσας εκτιμώμενης τιμής Q και μιας πιο ενημερωμένης εκτίμησης (Watkins & Dayan, 1992; Clifton & Laber, 2020).

Μια από τις πιο σημαντικές ιδιότητες του Q-learning είναι η εγγύηση σύγκλισής του. Οι Watkins & Dayan (1992) απέδειξαν λεπτομερώς ένα θεώρημα σύγκλισης, το οποίο βασίζεται σε αυτό που περιγράφηκε από τον Watkins (1989). Έδειξαν ότι ο Q-learning συγκλίνει στις βέλτιστες τιμές δράσης (Q^*) με πιθανότητα 1, εφόσον:

1. Όλες οι ενέργειες λαμβάνονται επανειλημμένα δειγματοληπτικά σε όλες τις καταστάσεις
2. Οι τιμές δράσης αναπαρίστανται διακριτά.
3. Ο ρυθμός μάθησης α ικανοποιεί τις συνθήκες Robbins-Monro (Clifton & Laber, 2020): $\sum_{t=0}^{\infty} \alpha_t = \infty$ και $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$.

Ο Q-learning έχει εφαρμογές σε ένα ευρύ φάσμα τομέων:

- **Προσωποποιημένη Ιατρική:** Στο πλαίσιο της προσωποποιημένης ιατρικής, ο Q-learning πεπερασμένου ορίζοντα (finite-horizon Q-learning) είναι το κύριο εργαλείο για την εκτίμηση βέλτιστων στρατηγικών θεραπείας, γνωστών ως πλάνα θεραπείας. Αυτό περιλαμβάνει προβλήματα όπως η επιλογή της καλύτερης θεραπείας για έναν ασθενή σε διαφορετικά στάδια της νόσου, λαμβάνοντας υπόψη την προηγούμενη ανταπόκριση.
- **Κινητή Υγεία (Mobile Health):** Ο Q-learning άπειρου ορίζοντα είναι ολοένα και πιο σημαντικός στον αναπτυσσόμενο τομέα της κινητής υγείας, όπου οι παρεμβάσεις πρέπει να προσαρμόζονται δυναμικά στη συμπεριφορά και τις ανάγκες των χρηστών.

- **Παιχνίδια:** Στην επιστήμη των υπολογιστών, οι μέθοδοι Q-learning έχουν επιτύχει αξιοσημείωτη απόδοση σε τομείς όπως το παιχνίδι, με χαρακτηριστικά παραδείγματα το AlphaGo.
- **Ρομποτική:** Ο Q-learning χρησιμοποιείται για να επιτρέψει στα ρομπότ να μάθουν βέλτιστες συμπεριφορές μέσω δοκιμής και λάθους.
- **Οικονομία:** Εφαρμογές σε χρηματοοικονομική μοντελοποίηση, βελτιστοποίηση χαρτοφυλακίου και στρατηγικές συναλλαγών.

Παρά την επιτυχία του, ο Q-learning αντιμετωπίζει προκλήσεις, ιδιαίτερα σε περιβάλλοντα με πολύ μεγάλους ή συνεχείς χώρους καταστάσεων και ενεργειών. Για την αντιμετώπιση αυτών των προκλήσεων, έχουν αναπτυχθεί διάφορες επεκτάσεις:

- **Q-learning με προσεγγίσεις συναρτήσεων (Approximation Functions):** Όταν οι καταστάσεις και οι ενέργειες είναι συνεχείς ή υπερβολικά πολλές για διακριτή αναπαράσταση, χρησιμοποιούνται προσεγγίσεις συναρτήσεων (π.χ., νευρωνικά δίκτυα) για την εκτίμηση των τιμών Q. Αυτό επιτρέπει τη γενίκευση σε μη ορατές καταστάσεις.
- **Deep Q-Networks (DQNs):** Συνδυάζουν Q-learning με βαθιά νευρωνικά δίκτυα, επιτρέποντας την επεξεργασία σύνθετων εισόδων (π.χ., εικόνες) και την επίλυση πολύπλοκων προβλημάτων, όπως αυτά που παρατηρούνται στο Atari (Clifton & Laber, 2020).
- **Περιβάλλοντα μηδενικής έκπτωσης αλλά απορροφητικά:** Οι Watkins & Dayan (1992) σκιαγράφησαν επεκτάσεις για αυτές τις περιπτώσεις, όπου η μάθηση μπορεί να συμβεί ακόμη και χωρίς συντελεστή έκπτωσης, εφόσον το περιβάλλον οδηγεί τελικά σε μια "απορροφητική" κατάσταση (π.χ., τέλος του παιχνιδιού).
- **Πολλαπλές αλλαγές τιμών Q ανά επανάληψη:** Οι Watkins & Dayan (1992) επίσης αναφέρουν επεκτάσεις όπου πολλές Q-τιμές μπορούν να ενημερωθούν σε κάθε επανάληψη, αντί για μία μόνο, γεγονός που μπορεί να επιταχύνει τη διαδικασία μάθησης.

Συνοψίζοντας, ο Q-learning αποτελεί ένα ισχυρό και ευέλικτο πλαίσιο για την εκμάθηση βέλτιστων πολιτικών σε δυναμικά περιβάλλοντα. Η απλότητά του ως επαυξητικός αλγόριθμος, σε συνδυασμό με τις ισχυρές θεωρητικές εγγυήσεις σύγκλισης, τον καθιστούν ένα θεμελιώδες εργαλείο τόσο στην έρευνα όσο και στις εφαρμογές της ενισχυτικής μάθησης (Watkins & Dayan, 1992; Clifton & Laber, 2020).

2.2.2 Deep Q-Networks

Η περιγραφή που ακολουθεί βασίζεται κυρίως στις προσεγγίσεις των Mnih et al. (2015), Roderick et al. (2017) και Fan et.al, (2020), οι οποίοι παρουσίασαν και ανέλυσαν διεξοδικά τη μέθοδο DQN.

Τα βαθιά νευρωνικά δίκτυα (Deep Q-Networks – DQN) αποτελούν μια επέκταση του κλασικού αλγόριθμου Q-learning που χρησιμοποιεί βαθιά νευρωνικά δίκτυα για την προσέγγιση της συνάρτησης αξίας-ενέργειας. Αυτή η προσέγγιση έχει οδηγήσει σε μεγάλη επιτυχία σε εμπειρικό επίπεδο στην βαθιά ενισχυτική μάθηση, ωστόσο η θεωρητική της βάση είναι λιγότερο κατανοητή (Fan et.al, 2020).

Βασικές Συνεισφορές του DQN

Ο αλγόριθμος DQN, όπως προτάθηκε από τους Mnih et al(2015), έχει τρεις κύριες συνεισφορές:

1. **Αρχιτεκτονική βαθύς συνελκτικού νευρωνικού δικτύου (deep convolutional neural net architecture):** Χρησιμοποιείται για την προσέγγιση της συνάρτησης Q. Αυτή η αρχιτεκτονική παρέχει έναν γενικό μηχανισμό για την εκτίμηση των τιμών της συνάρτησης Q από ένα σύντομο ιστορικό καρέ εικόνων (συγκεκριμένα, τα τελευταία 4 καρέ εμπειρίας) (Roderick et.al, 2017).
2. **Χρήση mini-batches τυχαίων δεδομένων εκπαίδευσης:** Αντί για ενημερώσεις ενός βήματος στην τελευταία εμπειρία, ο DQN χρησιμοποιεί mini-batches τυχαίων δειγμάτων. Αυτό επιτυγχάνεται μέσω της εμπειρίας επανάληψης (experience replay), όπου μια μνήμη αναπαραγωγής εμπειριών αποθηκεύει την τροχιά της Markov Decision Process (MDP). Σε κάθε επανάληψη του DQN, ένα mini-batch από καταστάσεις, ενέργειες, ανταμοιβές και επόμενες καταστάσεις λαμβάνονται δειγματοληπτικά από τη μνήμη αναπαραγωγής εμπειριών επανάληψης για την εκπαίδευση του δικτύου Q. Ο σκοπός της εμπειρίας επανάληψης είναι να επιτευχθεί σταθερότητα διακόπτοντας την χρονική εξάρτηση μεταξύ των παρατηρήσεων που χρησιμοποιούνται στην εκπαίδευση του βαθύς νευρωνικού δικτύου (Roderick et.al, 2017), (Fan et.al, 2020).
3. **Χρήση παλαιότερων παραμέτρων δικτύου:** Χρησιμοποιούνται για την εκτίμηση των Q-τιμών της επόμενης κατάστασης. Το target network συγχρονίζεται με το Q-network μετά από κάθε περίοδο επαναλήψεων, οδηγώντας σε μια σύζευξη μεταξύ των δύο δικτύων. Αυτή η προσέγγιση παρέχει έναν σταθερό στόχο εκπαίδευσης για τη συνάρτηση του δικτύου (Roderick et.al, 2017), (Fan et.al, 2020).

Προκλήσεις και Λύσεις

Παρά την επιτυχία του, η θεωρητική ανάλυση του DQN είναι εξαιρετικά δύσκολη λόγω των ακόλουθων διαφορών από το κλασικό Q-learning(Fan et.al, 2020):

- **Αστάθεια με προσέγγιση συνάρτησης:** Σε αλγόριθμους ενισχυτικής μάθησης χρονικής διαφοράς βασισμένους σε online gradients, η προσέγγιση της συνάρτησης αξίας-ενέργειας οδηγεί συχνά σε αστάθεια, ακόμη και με γραμμική προσέγγιση συνάρτησης. Η εμπειρία επανάληψης είναι η βασική τεχνική που χρησιμοποιείται στο DQN για την επίτευξη σταθερότητας (Roderick et.al, 2017), (Fan et.al, 2020).
- **Εκπαίδευση νευρωνικού δικτύου:** Ακόμη και αν διορθώσουμε το target network και επικεντρωθούμε στην ενημέρωση του Q-network, το υποπρόβλημα της εκπαίδευσης ενός νευρωνικού δικτύου παραμένει θεωρητικά λιγότερο κατανοητό (Fan et.al, 2020).

Θεωρητική Ανάλυση και Βελτιώσεις

Υπό ήπιες παραδοχές, έχουν καθοριστεί οι ρυθμοί σύγκλισης του αλγορίθμου και των στατιστικών για τις συναρτήσεις αξίας-ενέργειας της επαναληπτικής ακολουθίας πολιτικής που λαμβάνεται από τον DQN. Συγκεκριμένα, το στατιστικό σφάλμα χαρακτηρίζει τη μεροληψία και τη διασπορά που προκύπτουν από την προσέγγιση της συνάρτησης αξίας-ενέργειας χρησιμοποιώντας βαθύ νευρωνικό δίκτυο, ενώ το αλγοριθμικό σφάλμα συγκλίνει στο μηδέν με γεωμετρικό ρυθμό (Fan et.al, 2020).

Πρακτικές Λεπτομέρειες Υλοποίησης

Η αναπαραγωγή των αποτελεσμάτων του DQN μπορεί να είναι δύσκολη, καθώς οι αρχικές επιστημονικές δημοσιεύσεις δεν περιγράφουν πάντα λεπτομερώς κάθε σημαντική ρύθμιση παραμέτρων και λύση μηχανικής λογισμικού. Κρίσιμες λεπτομέρειες υλοποίησης περιλαμβάνουν (Roderick et.al, 2017):

- **Αρχικοποίηση επεισοδίων:** Κάθε επεισόδιο ξεκινά με έναν τυχαίο αριθμό "No-op" δράσεων χαμηλού επιπέδου Atari (μεταξύ 0 και 30) για να μετατοπιστούν τα καρέ που βλέπει ο πράκτορας.
- **Πλήρωση μνήμης εμπειρίας:** Πριν από οποιαδήποτε βήματα gradient descent, μια τυχαία πολιτική εκτελείται για 50.000 βήματα για να γεμίσει η μνήμη με εμπειρίες, αποφεύγοντας την υπερ-προσαρμογή σε πρώιμες εμπειρίες.
- **Συχνότητα ενημέρωσης δικτύου:** Η αρχική υλοποίηση του DQN λαμβάνει ένα βήμα gradient descent μόνο σε κάθε 4 βήματα περιβάλλοντος του αλγορίθμου, αντί για κάθε βήμα. Αυτό αυξάνει σημαντικά την ταχύτητα εκπαίδευσης και κάνει τη μνήμη εμπειρίας να μοιάζει περισσότερο με την κατανομή κατάστασης της τρέχουσας πολιτικής.

- **Τερματισμός στην απώλεια ζωών:** Σε παιχνίδια με "ζωές", η απώλεια μιας ζωής θεωρείται ως μια κατάσταση τερματισμού στην MDP κατά τη διάρκεια της εκπαίδευσης. Αυτό βελτιώνει την πρόωμη εκπαίδευση και τη σταθερότητα, και σε πιο πολύπλοκα παιχνίδια, βελτιώνει σημαντικά τη συνολική απόδοση.
- **Βελτιστοποίηση Gradient Descent:** Η αρχική υλοποίηση του DQN χρησιμοποιεί μια παραλλαγή του αλγορίθμου RMSProp που περιλαμβάνει έναν παράγοντα ορμής, ο οποίος απαιτεί προσαρμογή του ρυθμού μάθησης.

Διακυμαινόμενη Απόδοση και Καταστροφική Λήθη

Είναι σημαντικό να σημειωθεί ότι η απόδοση του DQN δεν βελτιώνεται απαραίτητα σταθερά. Είναι σύνηθες να παρατηρείται "καταστροφική λήθη" (catastrophic forgetting), όπου η απόδοση του πράκτορα μπορεί να μειωθεί σημαντικά μετά από μια περίοδο μάθησης. Αυτό οφείλεται στην εγγενή αστάθεια της προσέγγισης της Q-συνάρτησης σε έναν μεγάλο χώρο καταστάσεων χρησιμοποιώντας ενημερώσεις Bellman. Για την αντιμετώπιση αυτού, προτείνεται η αποθήκευση των παραμέτρων του δικτύου που οδήγησαν στην καλύτερη απόδοση δοκιμής. Επιπλέον, το κλιπ του gradient του όρου σφάλματος (σε τιμές μεταξύ 1.0 και -1.0) βελτιώνει περαιτέρω τη σταθερότητα του αλγορίθμου, αποτρέποντας οποιαδήποτε μεμονωμένη ενημέρωση mini-batch από το να αλλάξει σημαντικά τις παραμέτρους (Roderick et.al, 2017).

Ένας άλλος λόγος για την καταστροφική λήθη είναι ότι ο αλγόριθμος μαθαίνει ένα proxy (τις Q-τιμές) για μια πολιτική αντί να προσεγγίζει την πολιτική απευθείας. Αυτό μπορεί να οδηγήσει σε ενημερώσεις μάθησης που αυξάνουν την ακρίβεια ενός Q-συναρτησιακού προσεγγιστή, ενώ μειώνουν την απόδοση της προκύπτουσας πολιτικής. Επιπλέον, πολύ μικρά σφάλματα στις Q-τιμές μπορούν να οδηγήσουν σε πολύ διαφορετικές πολιτικές, καθιστώντας δύσκολη την εκμάθηση μακροπρόθεσμων πολιτικών (Roderick et.al, 2017).

2.2.3 Αλγόριθμοι Κλίσης Πολιτικής (Policy Gradient Algorithms)

REINFORCE

Οι αλγόριθμοι REINFORCE αποτελούν μια γενική κατηγορία αλγορίθμων ενισχυτικής μάθησης για νευρωνικά δίκτυα με στοχαστικές μονάδες. Αυτοί οι αλγόριθμοι προσαρμόζουν τα βάρη των συνδέσεων σε μια κατεύθυνση που ακολουθεί την κλίση της αναμενόμενης ενίσχυσης, τόσο σε εργασίες άμεσης ενίσχυσης όσο και σε ορισμένες περιορισμένες μορφές εργασιών καθυστερημένης ενίσχυσης. Το σημαντικό χαρακτηριστικό τους είναι ότι το επιτυγχάνουν αυτό χωρίς να υπολογίζουν ρητά εκτιμήσεις κλίσης ή να αποθηκεύουν πληροφορίες από τις οποίες θα μπορούσαν να υπολογιστούν τέτοιες εκτιμήσεις (Williams, 1992).

Βασικές Αρχές και Χαρακτηριστικά:

- **Στοχαστικές Μονάδες:** Οι αλγόριθμοι REINFORCE λειτουργούν σε δίκτυα που περιέχουν στοχαστικές μονάδες, δηλαδή μονάδες των οποίων η έξοδος παράγεται τυχαία από μια συνάρτηση κατανομής (Williams, 1992).
- **Ενισχυτική Μάθηση και Βελτιστοποίηση Συναρτήσεων:** Κάθε μη-συνειρμικός αλγόριθμος ενισχυτικής μάθησης μπορεί να θεωρηθεί ως μια μέθοδος για την εκτέλεση βελτιστοποίησης συνάρτησης μέσω δειγματοληψίας (ενδεχομένως με θόρυβο) τιμών συνάρτησης. Στο πλαίσιο της βελτιστοποίησης συνάρτησης, η τιμή της συνάρτησης χρησιμοποιείται ως το σήμα ενίσχυσης που παρέχεται στο δίκτυο (Williams & Peng, 1991).
- **Κριτήριο Απόδοσης:** Το κριτήριο απόδοσης που βελτιστοποιείται είναι η αναμενόμενη τιμή του σήματος ενίσχυσης, υπό την προϋπόθεση μιας συγκεκριμένης επιλογής παραμέτρων του συστήματος μάθησης. Για ένα δίκτυο ενισχυτικής μάθησης, αυτό εκφράζεται ως $E\{r|W\}$, όπου E είναι ο συντελεστής προσδοκίας, r το σήμα ενίσχυσης και W ο πίνακας βαρών του δικτύου. Ο στόχος είναι η εύρεση του W όπου το $E\{r|W\}$ είναι μέγιστο (Williams, 1992).
- **Αρχή της Κλίσης (Gradient Following):** Ο αλγόριθμος "ανεβαίνει" στατιστικά την κλίση της αναμενόμενης ενίσχυσης. Αυτό σημαίνει ότι ο μέσος φορέας ενημέρωσης στο χώρο των βαρών βρίσκεται σε μια κατεύθυνση στην οποία το μέτρο απόδοσης αυξάνεται (Williams & Peng, 1991) (Williams, 1992).
- **Μη-μοντελοποιημένοι(Non-model-based):** Οι αλγόριθμοι REINFORCE δεν απαιτούν ρητό υπολογισμό ή αποθήκευση εκτιμήσεων κλίσης, γεγονός που τους καθιστά "απλούς" ή "μη-μοντελοποιημένους" (Williams, 1992).

Κανόνας Ενημέρωσης Βαρών:

Για ένα δίκτυο που εκτελεί μια συνειρμική εργασία άμεσης ενίσχυσης, τα βάρη προσαρμόζονται μετά τη λήψη της τιμής ενίσχυσης r σε κάθε δοκιμή. Ο αλγόριθμος REINFORCE ορίζει την προσαύξηση κάθε παραμέτρου w_{ij} στο δίκτυο με την ποσότητα (Williams, 1992):

$$\Delta w_{ij} = a_{ij}(r - b_{ij})e_{ij}$$

Όπου:

- a_{ij} είναι ο συντελεστής ρυθμού μάθησης, ο οποίος είναι μη αρνητικός και μπορεί να εξαρτάται από το w^j και το χρόνο t .

- b_{ij} είναι η "γραμμή βάσης ενίσχυσης" (reinforcement baseline), η οποία είναι υπό συνθήκες ανεξάρτητη της y_i , δεδομένου του W και του x^i
- $e_{ij} = \theta \ln g_i / \theta w_{ij}$ ονομάζεται "χαρακτηριστική επιλεξιμότητα" του w_{ij}

Το όνομα REINFORCE είναι ένα ακρωνύμιο για "REward Increment = Nonnegative Factor Offset Reinforcement Characteristic Eligibility". (Williams, 1992). Για τις μονάδες Bernoulli-λογιστικής (Bernoulli logistic units), οι οποίες είναι κατάλληλες για συναρτήσεις που ορίζονται σε δυαδικές n -άδες, ο κανόνας ενημέρωσης βαρών απλοποιείται:

$$\Delta w_{ij} = \alpha_{ij} (r - b_{ij}) (y_i - p_i) x_j$$

όπου p_i είναι η πιθανότητα η μονάδα i να επιλέξει 1 ως τιμή εξόδου της. Όταν $\alpha_{ij} = \alpha$ και

$b_{ij} = b$ για όλα τα i και j , και το b δεν εξαρτάται από την άμεσα λαμβανόμενη ενίσχυση r , τότε ισχύει:

$$E\{\Delta W | W\} = \alpha \nabla_w E\{r | W\}$$

Αυτό δείχνει ότι ο μέσος φορέας ενημέρωσης των βαρών είναι ανάλογος με την κλίση της αναμενόμενης ενίσχυσης (Williams & Peng, 1991) (Williams, 1992).

Μια αξιοσημείωτη παραλλαγή είναι ο αλγόριθμος REINFORCE/MENT. Αυτός ο αλγόριθμος συνδυάζει την προσέγγιση REINFORCE με τη μεγιστοποίηση της εντροπίας. Η χρήση της μεγιστοποίησης της εντροπίας έχει σχεδιαστεί για να διατηρεί ζωντανή την αναζήτηση, αποτρέποντας τη σύγκλιση σε μία μόνο επιλογή εξόδου, ειδικά όταν πολλές επιλογές οδηγούν σε περίπου την ίδια τιμή ενίσχυσης. Ο αλγόριθμος REINFORCE/MENT χρησιμοποιεί ένα σήμα ενίσχυσης που συνδυάζει την "εξωτερική" ενίσχυση (όπως η τιμή της συνάρτησης που βελτιστοποιείται) με μια "εσωτερική" ενίσχυση που επιβραβεύει την ποικιλία. Αυτή η εσωτερική συμβολή είναι ανάλογη με την εντροπία της κατανομής των διανυσμάτων εξόδου που παράγονται από το δίκτυο, κάτι που έχει ως αποτέλεσμα το δίκτυο να είναι πρόθυμο να θυσιάσει κάποια απόδοση για να επιτύχει υψηλότερη εντροπία και να συνεχίσει να εξερευνά (Williams & Peng, 1991).

Actor Critic

Οι αλγόριθμοι Actor-Critic αποτελούν μια από τις πιο δημοφιλείς κατηγορίες αλγορίθμων ενισχυτικής μάθησης και είναι ιδιαίτερα χρήσιμοι σε εφαρμογές πραγματικού κόσμου, όπως η ρομποτική, ο έλεγχος ισχύος και την οικονομία (Grondman et.al 2012).

Βασική Έννοια και Συστατικά

Στο πλαίσιο της ενισχυτικής μάθησης, ένας πράκτορας βελτιστοποιεί τη συμπεριφορά του αλληλεπιδρώντας με το περιβάλλον του. Μετά από κάθε ενέργεια που εκτελεί σε μια δεδομένη κατάσταση, λαμβάνει μια ανταμοιβή που του δείχνει την ποιότητα της ενέργειας.

Στόχος του πράκτορα είναι να βρει μια "πολιτική" που να μεγιστοποιεί τη συνολική συσσωρευμένη ανταμοιβή.

Οι αλγόριθμοι Actor-Critic συνδυάζουν δύο βασικά συστατικά:

- **Actor (Πράκτορας Πολιτικής):** Αναφέρεται στην πολιτική που καθορίζει τις ενέργειες που πρέπει να εκτελεστούν. Ο πράκτορας πολιτικής είναι υπεύθυνος για τη δημιουργία ενεργειών δεδομένης της τρέχουσας κατάστασης του περιβάλλοντος (Jia & Zhou, 2022).
- **Critic (Εκτιμητής Αξίας):** Αναφέρεται στη συνάρτηση αξίας (value function) που αξιολογεί την απόδοση μιας πολιτικής. Ο εκτιμητής αξίας κρίνει την επιλεγμένη πολιτική και καθοδηγεί τη βελτίωση του actor (Jia & Zhou, 2022).

Πλεονεκτήματα

Οι αλγόριθμοι ενισχυτικής μάθησης μπορούν να κατηγοριοποιηθούν σε τρεις τύπους: actor-only, critic-only και actor-critic.

- **Actor-only** μέθοδοι, ενώ επιτρέπουν τη δημιουργία ενός φάσματος συνεχών ενεργειών μέσω παραμετροποιημένης πολιτικής, υποφέρουν από υψηλή διακύμανση στις εκτιμήσεις της κλίσης, οδηγώντας σε αργή μάθηση (Grondman et.al 2012).
- **Critic-only** μέθοδοι, που χρησιμοποιούν μάθηση χρονικής διαφοράς (Temporal Difference - TD), έχουν χαμηλότερη διακύμανση στις εκτιμήσεις της αναμενόμενης ανταμοιβής. Ωστόσο, η εξαγωγή μιας πολιτικής σε αυτές τις μεθόδους απαιτεί μια διαδικασία βελτιστοποίησης σε κάθε κατάσταση, κάτι που μπορεί να είναι υπολογιστικά έντονο, ειδικά σε συνεχή χώρο ενεργειών. Συχνά, αυτό οδηγεί σε διακριτοποίηση του χώρου ενεργειών, υπονομεύοντας την ικανότητα χρήσης συνεχών ενεργειών και εύρεσης του πραγματικού βέλτιστου (Grondman et.al 2012) (Jia & Zhou, 2022).

Οι μέθοδοι Actor-Critic συνδυάζουν τα πλεονεκτήματα και των δύο προσεγγίσεων. Ενώ ο παραμετροποιημένος πράκτορας πολιτικής παρέχει τη δυνατότητα υπολογισμού συνεχών ενεργειών χωρίς την ανάγκη για διαδικασίες βελτιστοποίησης στη συνάρτηση αξίας, ο εκτιμητής αξίας παρέχει στον πράκτορα πολιτικής γνώση χαμηλής διακύμανσης για την απόδοση. Συγκεκριμένα, η εκτίμηση του εκτιμητή αξίας για την αναμενόμενη ανταμοιβή επιτρέπει στον πράκτορα πολιτικής να ενημερώνεται με κλίσεις που έχουν χαμηλότερη διακύμανση, επιταχύνοντας τη διαδικασία μάθησης. Ένα πλεονέκτημα των actor-critic αλγορίθμων είναι ότι επιτυγχάνουν ισχυρές εγγυήσεις σύγκλισης (Grondman et.al 2012), (Awate, 2002).

Τύποι Αλγορίθμων Actor-Critic

Υπάρχουν δύο κύριοι τύποι αλγορίθμων Actor-Critic, που διακρίνονται από τον τρόπο υπολογισμού της κλίσης της πολιτικής:

- **Standard Gradient Actor-Critic Algorithms:** Αυτοί οι αλγόριθμοι βασίζονται στην τυπική κλίση πολιτικής (Grondman et.al 2012).
- **Natural Gradient Actor-Critic Algorithms:** Τα τελευταία χρόνια, η "φυσική κλίση" έχει γίνει πιο δημοφιλής. Αυτοί οι αλγόριθμοι χρησιμοποιούν τη Fisher Information Matrix (FIM) για να υπολογίσουν την κλίση σε έναν χώρο κατανομών πιθανοτήτων, κάτι που μπορεί να οδηγήσει σε ταχύτερη σύγκλιση (Grondman et.al 2012).

Ενώ η πλειονότητα των αλγορίθμων RL έχει αναπτυχθεί για διακριτού χρόνου Markovian Decision Processes (MDPs), υπάρχει αυξανόμενο ενδιαφέρον για την ενισχυτική μάθηση σε συνεχή χρόνο και χώρο, καθώς ο κόσμος είναι εγγενώς συνεχούς χρόνου και οι εφαρμογές απαιτούν συχνά συνεχή αλληλεπίδραση (π.χ., διαπραγμάτευση μετοχών υψηλής συχνότητας, αυτόνομη οδήγηση, πλοήγηση ρομπότ). Νέες προσεγγίσεις αναπτύσσονται για να γεφυρώσουν αυτό το χάσμα, μετατρέποντας την εκτίμηση της κλίσης πολιτικής σε ένα πρόβλημα αξιολόγησης πολιτικής (policy evaluation - PE) χρησιμοποιώντας μαθηματικά εργαλεία όπως οι συνθήκες ορθογωνιότητας martingale. Αυτό επιτρέπει την ανάπτυξη model-free actor-critic αλγορίθμων που μπορούν να εφαρμοστούν σε προβλήματα με συνεχή δυναμική (Jia & Zhou, 2022).

Οι μέθοδοι Actor-Critic έχουν σημειώσει επιτυχία σε πολλές εφαρμογές του πραγματικού κόσμου, όπως το AlphaGo και η επιδέξια χειραγώγηση χεριών. Παραδείγματα εφαρμογών σε συνεχή χρόνο περιλαμβάνουν προβλήματα όπως το cart-pole swing-up, τη διαχείριση χαρτοφυλακίου, τον έλεγχο κυκλοφορίας και την αυτόνομη οδήγηση (Jia & Zhou, 2022).

2.2.4 Proximal Policy Optimization (PPO)

Ο αλγόριθμος Proximal Policy Optimization (PPO) αποτελεί μια οικογένεια μεθόδων πολιτικής κλίσης (policy gradient) για την ενισχυτική μάθηση. Πρόκειται για έναν από τους πιο επιτυχημένους αλγορίθμους βαθιάς ενισχυτικής μάθησης, επιτυγχάνοντας κορυφαία απόδοση σε ένα ευρύ φάσμα δύσκολων εργασιών, όπως η προσομοίωση ρομποτικής κίνησης και το παιχνίδι Atari (Schulman et al, 2017).

Βασική Λειτουργία και Χαρακτηριστικά

Ο PPO εναλλάσσεται μεταξύ της δειγματοληψίας δεδομένων μέσω αλληλεπίδρασης με το περιβάλλον και της βελτιστοποίησης μιας προσεγγιστικής συνάρτησης (surrogate objective function) χρησιμοποιώντας στοχαστική ανάβαση κλίσης (stochastic gradient ascent). Σε αντίθεση με τις τυπικές μεθόδους πολιτικής κλίσης που εκτελούν μία ενημέρωση κλίσης ανά δείγμα δεδομένων, ο PPO προτείνει μια νέα συνάρτηση στόχου που επιτρέπει πολλαπλές εποχές ενημερώσεων minibatch (Schulman et al, 2017).

Πλεονεκτήματα έναντι άλλων μεθόδων

- **Σύγκριση με TRPO:** Ο PPO έχει ορισμένα από τα πλεονεκτήματα του Trust Region Policy Optimization (TRPO), αλλά είναι πολύ πιο απλός στην υλοποίηση, πιο γενικός και έχει καλύτερη πολυπλοκότητα δείγματος (εμπειρικά). Ο TRPO χρησιμοποιεί έναν αυστηρό περιορισμό (hard constraint) στο μέγεθος της ενημέρωσης της πολιτικής μέσω της απόκλισης KL (Kullback-Leibler divergence). Ωστόσο, η πολύπλοκη βελτιστοποίηση δεύτερης τάξης που απαιτείται από ο TRPO τον καθιστά υπολογιστικά αναποτελεσματικό και δύσκολο να εφαρμοστεί σε προβλήματα μεγάλης κλίμακας με πολύπλοκες αρχιτεκτονικές δικτύων. Ο PPO μειώνει σημαντικά αυτή την πολυπλοκότητα υιοθετώντας έναν μηχανισμό "κοψίματος" (clipping mechanism) για να αποφύγει την επιβολή του αυστηρού περιορισμού, επιτρέποντας τη χρήση ενός βελτιστοποιητή πρώτης τάξης, όπως η μέθοδος Gradient Descent, για τη βελτιστοποίηση του στόχου (Schulman et al, 2017), (Wang et.al, 2020).
- **Στόχος Κοψίματος (Clipped Surrogate Objective):** Ένα βασικό στοιχείο του PPO είναι η χρήση μιας περικομμένης συνάρτησης στόχου, η οποία σχηματίζει μια απαισιόδοξη εκτίμηση (δηλαδή, ένα κάτω όριο) της απόδοσης της πολιτικής. Ο κύριος στόχος που προτείνεται είναι:

$$L^{CLIP}(\Theta) = E_t^{\epsilon}$$

όπου $r_t(\Theta) = \frac{p_{i_{\Theta}}(a_t \vee s_t)}{p_{i_{\Theta_{old}}}(a_t \vee s_t)}$ είναι ο λόγος πιθανότητας της νέας πολιτικής προς την παλιά.

Το ϵ είναι μια υπερπαραμέτρος (π.χ., $\epsilon=0.2$). Ο δεύτερος όρος τροποποιεί τη συνάρτηση στόχου περικόπτοντας τον λόγο πιθανότητας, γεγονός που αφαιρεί το κίνητρο να μετακινηθεί το r_t εκτός του διαστήματος $[1-\epsilon, 1+\epsilon]$.

Η λήψη του ελαχίστου (min) μεταξύ του "κομμένου" και του "μη κομμένου" στόχου διασφαλίζει ότι ο τελικός στόχος είναι ένα κάτω όριο (μια απαισιόδοξη εκτίμηση) του μη κομμένου στόχου. Αυτό σημαίνει ότι η αλλαγή στον λόγο πιθανότητας αγνοείται μόνο όταν βελτιώνει τον στόχο, αλλά συμπεριλαμβάνεται όταν τον χειροτερεύει (Schulman et al, 2017).

Απόδοση και Εφαρμογές

Ο PPO υπερέρχει άλλων online μεθόδων πολιτικής κλίσης. Σε εργασίες συνεχούς ελέγχου, αποδίδει καλύτερα από τους αλγορίθμους με τους οποίους συγκρίνεται. Στο Atari, αποδίδει σημαντικά καλύτερα (όσον αφορά την πολυπλοκότητα δείγματος) από τον A2C και παρόμοια με τον ACER, παρόλο που είναι πολύ πιο απλός. Συνολικά, ο PPO επιτυγχάνει μια ευνοϊκή ισορροπία μεταξύ της πολυπλοκότητας δείγματος, της απλότητας και του χρόνου εκτέλεσης (Schulman et al, 2017)

Προκλήσεις και Βελτιώσεις

Παρά την επιτυχία του, η βελτιστοποιητική συμπεριφορά του PPO δεν είναι ακόμα πλήρως κατανοητή. Έχουν εκφραστεί ανησυχίες σχετικά με το αν ο PPO μπορεί να περιορίσει αυστηρά τον λόγο πιθανότητας ή να επιβάλει έναν καλά καθορισμένο περιορισμό "περιοχής αξιοπιστίας" (trust region constraint), κάτι που μπορεί να οδηγήσει σε αστάθεια στην απόδοση. Αυτό οφείλεται κυρίως στο ότι ο PPO δεν μπορεί να αφαιρέσει πλήρως το κίνητρο για την απομάκρυνση της πολιτικής και στην εγγενή διαφορά μεταξύ των περιορισμών που υιοθετούνται από τον PPO και τον TRPO αντίστοιχα (Wang et.al, 2020)

Για να αντιμετωπιστούν αυτά τα ζητήματα, έχει προταθεί μια βελτιωμένη μέθοδος PPO, ονομαζόμενη Trust Region-based PPO with Rollback (TR-PPO-RB). Αυτή η μέθοδος υιοθετεί μια νέα συνάρτηση κοψίματος (clipping function) που υποστηρίζει μια συμπεριφορά "επιστροφής" (rollback) για να περιορίσει τον λόγο μεταξύ της νέας και της παλιάς πολιτικής. Επιπλέον, η συνθήκη ενεργοποίησης για το κόψιμο αντικαθίσταται από μια βασισμένη σε περιοχή αξιοπιστίας, η οποία είναι θεωρητικά αιτιολογημένη σύμφωνα με το θεώρημα της περιοχής αξιοπιστίας. Ο αλγόριθμος TR-PPO-RB(Trust Region Proximal Policy Optimization with Replay Buffer) συνδυάζει τα πλεονεκτήματα του TRPO και του PPO, όντας θεωρητικά αιτιολογημένος και απλός στην υλοποίηση με βελτιστοποίηση πρώτης τάξης. Τα υπολογιστικά αποτελέσματα δείχνουν ότι οι προτεινόμενοι αλγόριθμοι υπερτερούν των υφιστάμενων στις εργασίες Garnet(Wang et.al, 2020).

2.3 Πεδία Εφαρμογής Ενισχυτικής Μάθησης

Η Ενισχυτική Μάθηση είναι ένα δυναμικά αναπτυσσόμενο πεδίο της μηχανικής μάθησης που βρίσκει εφαρμογή σε πολλούς τομείς, από τη ρομποτική μέχρι την ιατρική και τα χρηματοοικονομικά. Ακολουθεί μια παρουσίαση των βασικών πεδίων εφαρμογής, με τεκμηρίωση από σχετικές επιστημονικές εργασίες.

Κύρια Πεδία Εφαρμογής Ενισχυτικής Μάθησης

1. **Ρομποτική:** Χρησιμοποιείται για να διδάξει σε ρομπότ πολύπλοκες ακολουθίες ενεργειών, όπως ο χειρισμός αντικειμένων και η πλοήγηση σε μεταβαλλόμενα περιβάλλοντα. Η μάθηση μέσω δοκιμής και σφάλματος είναι ιδανική για συστήματα που αλληλεπιδρούν συνεχώς με το περιβάλλον (Shimpi, 2025), (Ribeiro, 1999).
2. **Υγειονομική Περίθαλψη:** Χρησιμοποιείται για τη δημιουργία εξατομικευμένων θεραπευτικών πλάνων, για τη διαχείριση δόσεων φαρμάκων, καθώς και για την ανάλυση ιατρικών εικόνων (Shimpi, 2025).
3. **Χρηματοοικονομικά και Οικονομία:** Εφαρμογές περιλαμβάνουν την αυτόματη διαχείριση επενδυτικών χαρτοφυλακίων, τη λήψη αποφάσεων σε πραγματικό χρόνο στις αγορές, και την ανάπτυξη στρατηγικών συναλλαγών (Shimpi, 2025).

4. **Gaming και Τεχνητοί Παίκτες:** Χρησιμοποιείται εκτεταμένα σε παιχνίδια (π.χ. Atari, Go, Poker) για την εκπαίδευση πρακτόρων που ανταγωνίζονται ή υπερτερούν των ανθρώπων (Subramanian et al., 2020).
5. **Αθλητισμός και Καταγραφή Κίνησης:** Βοηθά στην ανάπτυξη προσομοιώσεων ανθρώπινης κίνησης (π.χ. σκι, γυμναστική) και στη βελτιστοποίηση κινήσεων σε αθλήματα και animation (Ashley, 2020).
6. **Αυτόνομη Οδήγηση και Πλοήγηση:** Συνεισφέρει στην εκμάθηση στρατηγικών οδήγησης, αποφάσεων σε συνθήκες κυκλοφορίας, και βελτιστοποίησης διαδρομών για αυτόνομα οχήματα (Diederichs, 2019).
7. **Ψυχολογία και Νευροεπιστήμες:** Χρησιμοποιείται ως υπολογιστικό μοντέλο για την κατανόηση της μάθησης και της λήψης αποφάσεων στον εγκέφαλο (Subramanian et al., 2020).
8. **Βιομηχανικές Διεργασίες και Έξυπνα Συστήματα:** Χρήση της ενισχυτικής μάθησης σε αυτόματους ελεγκτές, έξυπνη συντήρηση εξοπλισμού και αυτοματισμό παραγωγής σε βιομηχανικά περιβάλλοντα (Ribeiro, 1999).

Συνοψίζοντας, η ενισχυτική μάθηση έχει τεράστιο εύρος εφαρμογών σε τομείς όπου απαιτείται συνεχής αλληλεπίδραση με ένα δυναμικό περιβάλλον και λήψη αποφάσεων βάσει ενισχύσεων. Η πολυπλοκότητα και η ευελιξία της την καθιστούν μία από τις πιο υποσχόμενες τεχνολογίες της τεχνητής νοημοσύνης.

ΚΕΦΑΛΑΙΟ 3: ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΟΣ ΚΑΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΕΦΑΡΜΟΓΗΣ

Στο παρόν κεφάλαιο παρουσιάζεται η αρχιτεκτονική της εφαρμογής και τα βασικά χαρακτηριστικά του περιβάλλοντος προσομοίωσης που χρησιμοποιήθηκε για την εκπαίδευση του πράκτορα. Το περιβάλλον έχει τη μορφή ενός παιχνιδιού τύπου Endless Runner, όπου ο πράκτορας καλείται να αποφεύγει εμπόδια σε πραγματικό χρόνο, μεγιστοποιώντας τον χρόνο επιβίωσης και το σκορ. Το σενάριο αυτό χαρακτηρίζεται από:

- Συνεχή ροή εμποδίων που εμφανίζονται δυναμικά
- Περιορισμένο χώρο ενεργειών, με κύρια ενέργεια το άλμα
- Απαιτήσεις αντίληψης σε πραγματικό χρόνο ώστε ο πράκτορας να προσαρμόζεται άμεσα στις αλλαγές του περιβάλλοντος

Αυτά τα χαρακτηριστικά καθιστούν το περιβάλλον κατάλληλο για την εκπαίδευση μέσω τεχνικών ενισχυτικής μάθησης, ενώ η τελική επιλογή του αλγορίθμου εκπαίδευσης και οι λεπτομέρειες της υλοποίησης παρουσιάζονται στην ενότητα 3.4.

Αρχικά παρουσιάζεται η λογική σχεδίασης του περιβάλλοντος και οι βασικοί κανόνες λειτουργίας του. Ακολουθεί η ανάλυση των επιμέρους συστατικών που συνθέτουν τη σκηνή, καθώς και της μεταξύ τους αλληλεπίδρασης. Τέλος, γίνεται αναφορά στις ενέργειες και παρατηρήσεις του πράκτορα, στο σύστημα ανταμοιβών/ποινών και στη ρύθμιση των παραμέτρων εκπαίδευσης.

3.1 Περιγραφή Περιβάλλοντος Επίδειξης

Το σενάριο επίδειξης βασίζεται στη λογική ενός κλασικού side-scrolling παιχνιδιού επιβίωσης, όπου ο πράκτορας πρέπει να αποφεύγει διαδοχικά εμπόδια εκτελώντας άλματα με σωστό συγχρονισμό. Αν και ο πράκτορας εμφανίζεται να κινείται εμπρός μέσω animation, στην πραγματικότητα παραμένει ακίνητος και τα εμπόδια κινούνται προς το μέρος του.

Τα αντικείμενα που λειτουργούν ως εμπόδια (όπως φράχτες, βαρέλια και βράχια) δημιουργούνται περιοδικά και κινούνται με σταθερή ταχύτητα από προκαθορισμένο σημείο. Η συνεχής ροή τους αυξάνει προοδευτικά τη δυσκολία, απαιτώντας από τον πράκτορα αυξανόμενη ακρίβεια στις αντιδράσεις του.

Η κάμερα προβάλλει τη σκηνή από πλάγια θέση και παραμένει σταθερή καθ' όλη τη διάρκεια του επεισοδίου, παρέχοντας σαφή οπτική γωνία του διαθέσιμου χώρου δράσης. Ο πράκτορας έχει μόνο μία διαθέσιμη ενέργεια, το άλμα, και κάθε επεισόδιο ολοκληρώνεται όταν υπάρξει σύγκρουση με εμπόδιο.

Η απόδοσή του μετράται με σύστημα βαθμολόγησης, για κάθε επιτυχές άλμα που αποφεύγει εμπόδιο, προστίθενται 10 πόντοι. Το σκορ προβάλλεται σε πραγματικό χρόνο μέσω του περιβάλλοντος χρήστη (UI), και μετά από κάθε αποτυχία, το περιβάλλον επανεκκινείται για νέο επεισόδιο.

3.2 Δομή Σκηνής

Η σκηνή του παιχνιδιού οργανώνεται σε διακριτά δομικά στοιχεία, καθένα από τα οποία επιτελεί συγκεκριμένο ρόλο στη λειτουργία του συστήματος. Η δομή είναι ιεραρχική, επιτρέποντας ευελιξία τόσο στον σχεδιασμό όσο και στη διαχείριση του περιβάλλοντος.

Κατά την έναρξη της εφαρμογής, ο χρήστης μεταφέρεται αρχικά σε ένα απλό γραφικό μενού εκκίνησης, το οποίο περιλαμβάνει τα κουμπιά "Start Agent" και "Exit". Μέσω αυτού, παρέχεται η δυνατότητα έναρξης ενός νέου επεισοδίου ή εξόδου από την εφαρμογή(Σχήμα 3.1). Το μενού αυτό συμβάλλει στη δομημένη παρουσίαση της εμπειρίας χρήσης και διευκολύνει τον χειρισμό της εφαρμογής εκτός του περιβάλλοντος ανάπτυξης.

(Σχήμα 3.1, Αρχικό μενού εφαρμογής)

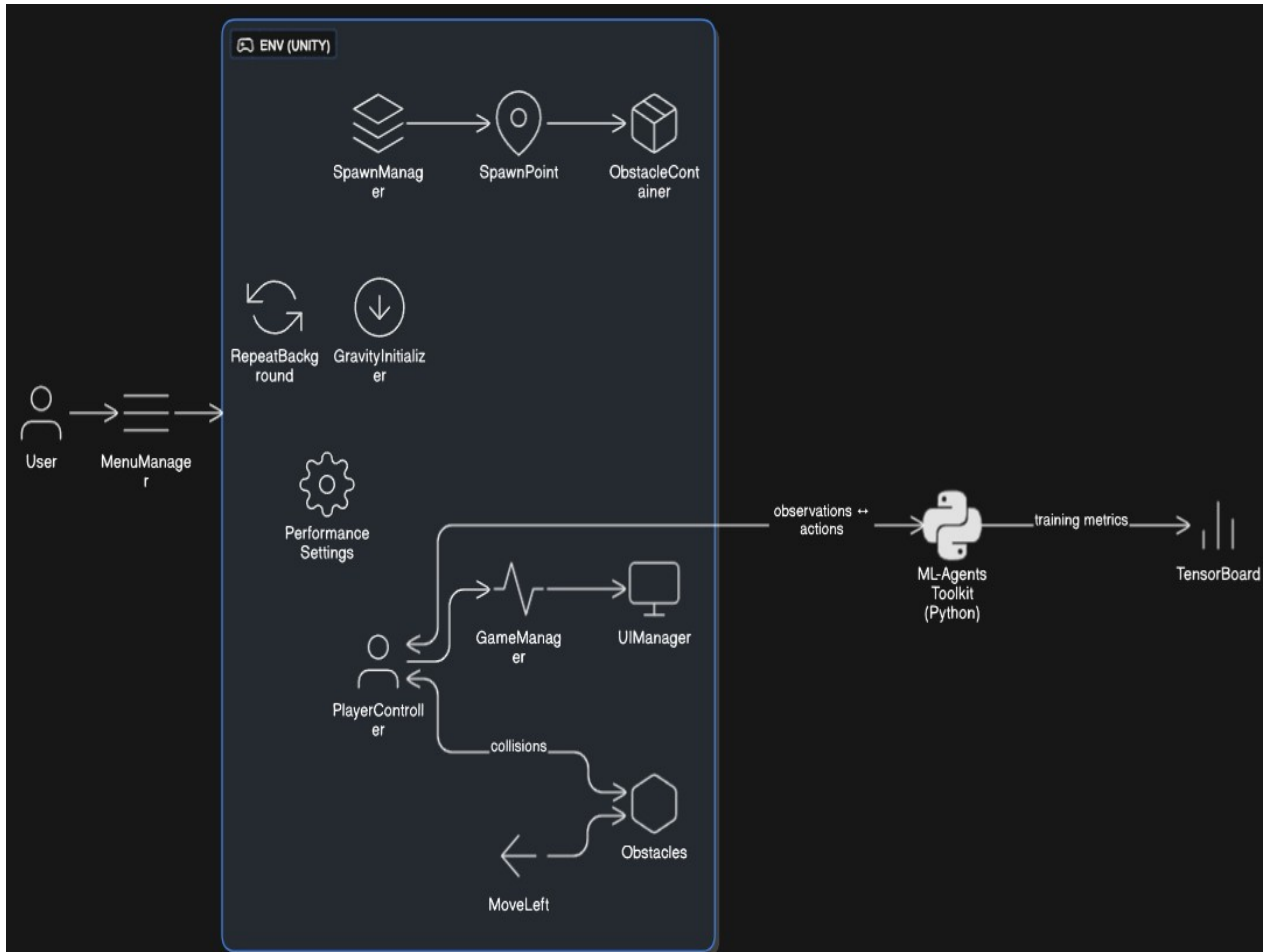


Ο βασικός κόμβος της σκηνής είναι το αντικείμενο Env, το οποίο περιλαμβάνει όλα τα αντικείμενα παιχνιδιού (Game Objects) που σχετίζονται με τη λειτουργία της εκπαίδευσης. Σε αυτό περιλαμβάνονται(Σχήμα 3.2):

- **Πράκτορας**, ένας ανθρωποειδής χαρακτήρας εξοπλισμένος με animation τρεξίματος, άλματος και αποτυχίας.
- **SpawnManager**, υπεύθυνος για τη δημιουργία των εμποδίων σε καθορισμένα χρονικά διαστήματα.
- **GameManager**, ο οποίος διαχειρίζεται τη ροή του παιχνιδιού και τον έλεγχο των επεισοδίων.

- **ObstacleContainer**, που φιλοξενεί δυναμικά όλα τα ενεργά εμπόδια, διευκολύνοντας τη διαχείρισή τους.
- **SpawnPoint**, σταθερό σημείο από όπου ξεκινά η κίνηση κάθε εμποδίου.

(Σχήμα 3.2, Διάγραμμα Αρχιτεκτονικής Συστήματος)



Τα εμπόδια έχουν υλοποιηθεί ως prefabs, επιτρέποντας την αναπαραγωγή, παραμετροποίηση και καταστροφή τους κατά τη διάρκεια του παιχνιδιού χωρίς να απαιτείται χειροκίνητη παρεμβολή. Η αναπαραγωγή γίνεται περιοδικά με τη μέθοδο `InvokeRepeating`.

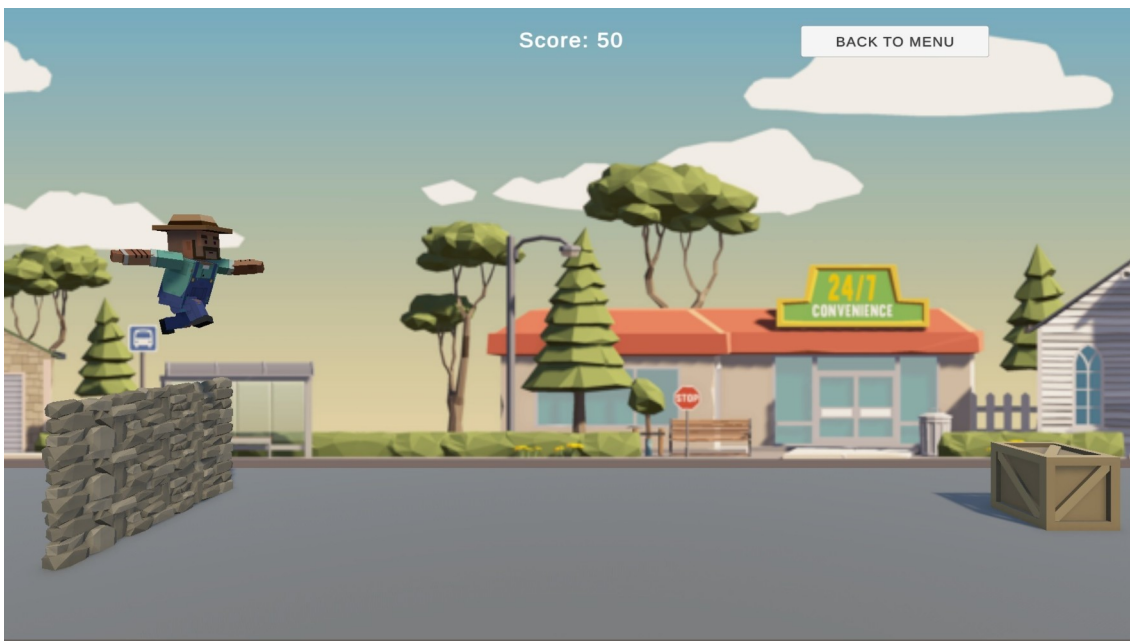
Ο πράκτορας είναι εξοπλισμένος με δύο Ray Perception Sensors (3D), οι οποίοι σαρώνουν το χώρο εμπρός του σε δύο διαφορετικές γωνίες. Η μία δέσμη κοιτάζει ευθεία, ενώ η άλλη ελαφρώς προς το έδαφος, ώστε να διευκολύνεται η αναγνώριση εμποδίων ακόμη και όταν αυτά βρίσκονται σε χαμηλότερο υψόμετρο κατά τη διάρκεια άλματος. Οι πληροφορίες που συλλέγονται μετατρέπονται σε παρατηρήσεις οι οποίες εισάγονται στο νευρωνικό δίκτυο.

Επιπλέον, το σύστημα περιλαμβάνει ένα Πλαίσιο (Canvas), το οποίο υποστηρίζεται από το EventSystem για διαχείριση του Περιβάλλοντος Χρήστη (User Interface - UI). Το γραφικό περιβάλλον περιλαμβάνει τα εξής στοιχεία:

- ScoreText, που προβάλλει το συνολικό σκορ του πράκτορα.
- MenuButton, για την επιστροφή στο κύριο μενού της εφαρμογής.

Η κάμερα έχει τοποθετηθεί σε σταθερή θέση πλάγια του πράκτορα, διατηρώντας αμετάβλητη την οπτική καθ' όλη τη διάρκεια ενός επεισοδίου. Ο φωτισμός παρέχεται από ένα Directional Light, χωρίς πρόσθετες δυναμικές σκιές ή εφέ που θα μπορούσαν να επηρεάσουν την απόδοση ή τη σταθερότητα του περιβάλλοντος(Σχήμα 3.3).

(Σχήμα 3.3, Gameplay παιχνιδιού)



3.3 Αλληλεπίδραση Πράκτορα – Περιβάλλοντος

Η συμπεριφορά του πράκτορα καθορίζεται από τον τρόπο με τον οποίο αντιλαμβάνεται το περιβάλλον και επιλέγει ενέργειες με βάση τις παρατηρήσεις του. Στο παρόν σύστημα, η αντίληψη επιτυγχάνεται μέσω δύο Ray Perception Sensors 3D, οι οποίοι ανιχνεύουν αντικείμενα εντός του οπτικού τους πεδίου και μετατρέπουν την πληροφορία αυτή σε αριθμητικά δεδομένα εισόδου για το μοντέλο ενισχυτικής μάθησης.

Ο ένας αισθητήρας κοιτά ευθεία εμπρός, ενώ ο άλλος είναι στραμμένος ελαφρώς προς τα κάτω, έτσι ώστε κατά τη διάρκεια άλματος να συνεχίζεται η ανίχνευση εμποδίων που βρίσκονται χαμηλά ή πλησιάζουν τη βάση του πράκτορα. Αυτή η επιλογή ενισχύει τη χωρική

κατανόηση της σκηνής και επιτρέπει πιο ακριβείς αποφάσεις σχετικά με το πότε είναι κατάλληλο να εκτελεστεί άλμα.

Η δράση που έχει στη διάθεσή του ο πράκτορας είναι μονοσήμαντη: η εκτέλεση άλματος. Δεν υπάρχει δυνατότητα διπλού άλματος ή άλλου τύπου κίνησης. Το γεγονός αυτό καθιστά το πρόβλημα απλούστερο από πλευράς action space, αλλά αυξάνει τις απαιτήσεις χρονισμού, καθώς κάθε ενέργεια πρέπει να εκτελείται με ακρίβεια και μόνο όταν είναι πραγματικά απαραίτητη.

Ο σχεδιασμός του μηχανισμού ανταμοιβής/ποινής ακολουθεί στρατηγική ενίσχυσης ορθής συμπεριφοράς και αποθάρρυνσης επαναλαμβανόμενων, τυχαίων ενεργειών. Συγκεκριμένα:

- Κάθε επιτυχές άλμα που αποφεύγει εμπόδιο αποδίδει +2.0 μονάδες ανταμοιβής
- Η σύγκρουση με εμπόδιο οδηγεί σε ποινή -0.8, ενώ τερματίζει και το επεισόδιο.
- Κάθε άλμα, ανεξαρτήτως αποτελέσματος, επιβαρύνεται με μικρή ποινή -0.06, ώστε να αποθαρρύνεται η υπερβολική ή τυχαία χρήση του.

Ο παραπάνω σχεδιασμός συμβάλλει στην εκμάθηση συντηρητικής και ακριβούς συμπεριφοράς. Ο πράκτορας τείνει να εκτελεί άλματα μόνο όταν η πρόβλεψη σύγκρουσης είναι υψηλή και η ανταμοιβή αναμένεται θετική. Αυτό ενισχύει την αποδοτικότητα και οδηγεί σε σταθερή βελτίωση της απόδοσής του με την πάροδο των επεισοδίων.

3.4 Επιλογή Αλγορίθμου PPO και Εκπαίδευση Πράκτορα

Βάσει των χαρακτηριστικών του περιβάλλοντος που περιγράφηκαν παραπάνω, επιλέχθηκε για την εκπαίδευση του πράκτορα ο αλγόριθμος Proximal Policy Optimization (PPO). Το συγκεκριμένο περιβάλλον τύπου Endless Runner απαιτεί συνεχή αντίληψη του χώρου, άμεση απόκριση σε δυναμικά εμπόδια και λήψη αποφάσεων σε πραγματικό χρόνο με περιορισμένο σύνολο ενεργειών (κυρίως το άλμα). Αυτές οι απαιτήσεις καθιστούν αναγκαίο έναν αλγόριθμο που να συνδυάζει σταθερότητα στη μάθηση και αποτελεσματικότητα στην προσαρμογή, χαρακτηριστικά που προσφέρει ο PPO. Η μεθοδολογία του επιτρέπει ελεγχόμενη διερεύνηση και προοδευτική βελτίωση της πολιτικής, αποφεύγοντας απότομες μεταβολές που οδηγούν σε αστάθεια.

Η παραμετροποίηση της εκπαίδευσης περιλαμβάνει, μεταξύ άλλων, τις εξής βασικές ρυθμίσεις:

- batch_size: 1024
- buffer_size: 20480

- learning_rate: 0.0003
- beta: 0.005
- epsilon: 0.2
- num_epoch: 3
- hidden_units: 128 σε 2 πλήρως συνδεδεμένα στρώματα
- discount factor (γ): 0.99
- time_horizon: 64
- max_steps: 2.500.000
- time_scale: 4.0 (για επιτάχυνση της προσομοίωσης)

Η εκπαίδευση πραγματοποιήθηκε εντός του Unity Editor, επιτρέποντας την άμεση παρακολούθηση της συμπεριφοράς του πράκτορα και τον έλεγχο της ορθότητας των παρατηρήσεων, ενεργειών και ανταμοιβών κατά τη διάρκεια των επεισοδίων. Αρχικά, πραγματοποιήθηκαν δοκιμές και επανασχεδιασμός του μηχανισμού επιβράβευσης/τιμωρίας, προκειμένου να αποφευχθούν φαινόμενα όπως η υπερβολική εκτέλεση αλμάτων ή η πλήρης απραξία.

Αφού επιτεύχθηκε σταθερή συμπεριφορά, η εκπαίδευση εξελίχθηκε ομαλά και ο πράκτορας εμφάνισε σταδιακή βελτίωση στην απόδοσή του. Λόγω της φύσης του προβλήματος όπου απαιτείται ακριβής χρονισμός και προσαρμογή σε μη προβλέψιμα μοτίβα εμποδίων χρειάστηκαν πολλές χιλιάδες βήματα έως ότου το δίκτυο κατορθώσει να γενικεύσει ικανοποιητικά τη στρατηγική αποφυγής.

ΚΕΦΑΛΑΙΟ 4: ΠΕΡΙΓΡΑΦΗ ΥΛΟΠΟΙΗΣΗΣ

Το παρόν κεφάλαιο εστιάζει στην υλοποίηση του διαδραστικού συστήματος που παρουσιάστηκε παραπάνω. Αφού έχει περιγραφεί αναλυτικά η θεωρητική βάση και η αρχιτεκτονική της εφαρμογής, ακολουθεί η ανάλυση των τεχνικών επιλογών, των εργαλείων που χρησιμοποιήθηκαν και των βημάτων που απαιτήθηκαν για την κατασκευή και εκπαίδευση του πράκτορα. Εξετάζονται επίσης οι προκλήσεις που αναδύθηκαν κατά τη διάρκεια της ανάπτυξης, καθώς και οι λύσεις που εφαρμόστηκαν για την επίτευξη σταθερής και αποδοτικής συμπεριφοράς.

4.1 Τεχνικές Επιλογές (Unity, ML-Agents, Python)

Για την ανάπτυξη και εκπαίδευση του πράκτορα αξιοποιήθηκε ένα σύνολο εργαλείων που εξασφάλισαν τη λειτουργικότητα του συστήματος:

- **Unity 3D** ως βασικό περιβάλλον ανάπτυξης και προσομοίωσης, για τη δημιουργία της σκηνής και τη διαχείριση των φυσικών αλληλεπιδράσεων.
- **ML-Agents Toolkit** για την ενσωμάτωση αλγορίθμων ενισχυτικής μάθησης, παρέχοντας έτοιμες κλάσεις διαχείρισης πρακτόρων και σύνδεσης με Python trainers
- **Python API & TensorBoard** για την εκτέλεση των εκπαιδύσεων και την οπτική παρακολούθηση των μετρικών απόδοσης.

Οι λεπτομέρειες εκπαίδευσης και οι ρυθμίσεις του PPO παρουσιάζονται στην ενότητα 3.4, ενώ η πρακτική διαδικασία εκπαίδευσης περιγράφεται αναλυτικά στην ενότητα 4.4.

4.2 Εφαρμογή Αρχιτεκτονικής σε Πρακτικό Επίπεδο

Η εφαρμογή της αρχιτεκτονικής του συστήματος πραγματοποιήθηκε μέσα από μια σειρά βημάτων που κάλυπταν τόσο την ανάπτυξη του περιβάλλοντος στο Unity όσο και τη διασύνδεση με τον αλγόριθμο ενισχυτικής μάθησης μέσω ML-Agents.

1. **Δημιουργία σκηνής στο Unity:** Αρχικά υλοποιήθηκε το περιβάλλον τύπου endless runner μέσα στο Unity 3D. Η σκηνή περιλάμβανε τον πράκτορα, τα εμπόδια, τον μηχανισμό δημιουργίας τους (Spawner Manager) και το σύστημα διαχείρισης του παιχνιδιού (Game Manager). Ο πράκτορας σχεδιάστηκε ως ένα ανθρωποειδές μοντέλο με animations τρεξίματος, άλματος και αποτυχίας, τα οποία συγχρονίζονταν με τις ενέργειες που προέκυπταν από το δίκτυο ενισχυτικής μάθησης.
2. **Ορισμός παρατηρήσεων και ενεργειών:** Για να είναι δυνατή η αλληλεπίδραση πράκτορα-περιβάλλοντος, καθορίστηκαν οι παρατηρήσεις που αντλούνται μέσω των Ray Perception Sensors (3D). Οι αισθητήρες τοποθετήθηκαν σε γωνίες που επιτρέπουν την αναγνώριση εμποδίων τόσο στο έδαφος όσο και σε χαμηλό ύψος κατά τη διάρκεια άλματος. Η ενέργεια του πράκτορα περιορίστηκε σε μία επιλογή, την εκτέλεση άλματος, καθιστώντας το πρόβλημα διακριτού χώρου ενεργειών.

3. **Σχεδιασμός συστήματος ανταμοιβής/ποινής:** Ο μηχανισμός ανταμοιβών ορίστηκε ώστε να κατευθύνει τη μαθησιακή διαδικασία ως εξής: +2.0 μονάδες για κάθε επιτυχή αποφυγή εμποδίου, -0.8 μονάδες και τερματισμός επεισοδίου σε περίπτωση σύγκρουσης και -0.06 μονάδες για κάθε άλμα ανεξάρτητα από την έκβαση, ώστε να αποφεύγεται η υπερβολική χρήση της ενέργειας. Αυτή η διαμόρφωση ενθάρρυνε τον πράκτορα να εκτελεί άλματα μόνο όταν ήταν απαραίτητο, οδηγώντας σε πιο συνεπή και αποδοτική στρατηγική.
4. **Διασύνδεση Unity με ML-Agents:** Για την εκπαίδευση χρησιμοποιήθηκε το πακέτο Unity ML-Agents Toolkit. Η συμπεριφορά του πράκτορα ορίστηκε σε ειδικό script (Agent Class), όπου υλοποιήθηκαν οι μέθοδοι OnEpisodeBegin, CollectObservations και OnActionReceived. Παράλληλα, δημιουργήθηκε το αρχείο παραμετροποίησης trainer_config.yaml, στο οποίο ορίστηκαν οι υπερπαραμέτροι του PPO (batch_size, buffer_size, learning_rate, γ κ.ά.).
5. **Εκτέλεση και παρακολούθηση εκπαίδευσης:** Η διαδικασία εκπαίδευσης πραγματοποιήθηκε μέσω Python API, με χρήση της εντολής mlagents-learn. Η πρόοδος του πράκτορα παρακολουθήθηκε μέσω TensorBoard, όπου καταγράφηκαν οι βασικές μετρικές (cumulative reward, policy loss, value loss). Παράλληλα, η προσομοίωση εκτελέστηκε σε αυξημένη ταχύτητα (time_scale = 4.0) ώστε να μειωθεί ο χρόνος εκπαίδευσης.
6. **Έλεγχος και βελτιστοποίηση συμπεριφοράς:** Κατά τη διάρκεια της ανάπτυξης εφαρμόστηκαν πολλαπλές δοκιμές με διαφορετικούς συνδυασμούς υπερπαραμέτρων και διαφορετικά reward functions, έως ότου επιτευχθεί σταθερή συμπεριφορά. Η επιλογή του τελικού checkpoint βασίστηκε στην καλύτερη συσσωρευμένη ανταμοιβή και στη συνεπή ικανότητα αποφυγής εμποδίων.

4.3 Προκλήσεις και Τεχνικές Λύσεις

Κατά τη διαδικασία ανάπτυξης και εκπαίδευσης του πράκτορα παρουσιάστηκαν αρκετές τεχνικές προκλήσεις, οι οποίες απαιτούσαν προσαρμογές τόσο στο επίπεδο σχεδίασης του περιβάλλοντος όσο και στη ρύθμιση των παραμέτρων εκπαίδευσης:

- **Αστάθεια στη Μάθηση:** Στα αρχικά στάδια παρατηρήθηκε έντονη διακύμανση στις τιμές της συσσωρευμένης ανταμοιβής, γεγονός που οδήγησε σε ασυνεπή συμπεριφορά του πράκτορα. Η λύση δόθηκε με αναθεώρηση του μηχανισμού ανταμοιβών, ώστε να επιβραβεύεται η στοχευμένη αποφυγή εμποδίων και να μειώνεται η τυχαία εκτέλεση ενεργειών.
- **Υπερεκπαίδευση (Overfitting):** Σε μακροχρόνιες περιόδους εκπαίδευσης ο πράκτορας παρουσίαζε policy degradation, με αποτέλεσμα απότομη πτώση της απόδοσης. Το πρόβλημα αντιμετωπίστηκε με περιορισμό του αριθμού βημάτων

εκπαίδευσης και επιλογή του καλύτερου checkpoint πριν την εμφάνιση φαινομένων αποδόμησης.

- **Αντιμετώπιση μη ρεαλιστικής συμπεριφοράς:** Σε ορισμένα πειράματα ο πράκτορας έτεινε να εκτελεί συνεχόμενα άλματα χωρίς λόγο. Η προσθήκη μικρής ποινής σε κάθε άλμα (-0.06) επέβαλε έναν μηχανισμό αυτορρύθμισης που αποθάρρυνε αυτή τη συμπεριφορά.
- **Διαχείριση χρόνου εκπαίδευσης:** Η εκπαίδευση ήταν χρονοβόρα λόγω του μεγάλου αριθμού επεισοδίων. Η χρήση της παραμέτρου `time_scale = 4.0` στο Unity επιτάχυνε σημαντικά την προσομοίωση, χωρίς να επηρεάσει την ποιότητα των αποτελεσμάτων.
- **Σταθερότητα Περιβάλλοντος:** Προβλήματα προέκυψαν από τη μη σωστή ανίχνευση συγκρούσεων, τα οποία οδηγούσαν σε λανθασμένη καταγραφή ανταμοιβών. Η βελτίωση του collider system στα εμπόδια και η ρύθμιση των Ray Perception Sensors εξάλειψαν το ζήτημα.

Οι παραπάνω τεχνικές λύσεις επέτρεψαν την ανάπτυξη ενός σταθερού και αξιόπιστου συστήματος μάθησης, βελτιώνοντας την ταχύτητα σύγκλισης και τη γενικότερη απόδοση του πράκτορα.

4.4 Διαδικασία Εκπαίδευσης

Η διαδικασία εκπαίδευσης του πράκτορα υλοποιήθηκε με βάση τις ρυθμίσεις PPO που αναφέρθηκαν στην ενότητα 3.4 και τα εργαλεία από την ενότητα 4.1. Η εκπαίδευση πραγματοποιήθηκε μέσα από το Unity Editor, επιτρέποντας την άμεση παρατήρηση της συμπεριφοράς του πράκτορα και τον έλεγχο της ορθότητας των παρατηρήσεων. Παράλληλα, το TensorBoard αξιοποιήθηκε για την παρακολούθηση κρίσιμων μετρικών (Cumulative reward, policy loss, value loss κλπ), προσφέροντας οπτική ανατροφοδότηση της πορείας της εκπαίδευσης.

Για τη βελτίωση της αξιοπιστίας χρησιμοποιήθηκαν checkpoints σε τακτά διαστήματα, τα οποία επέτρεπαν την επανεκκίνηση της διαδικασίας από ενδιάμεσα στάδια χωρίς απώλεια προόδου. Η εκπαίδευση συνεχίστηκε μέχρι την επίτευξη σταθερής συμπεριφοράς από τον πράκτορα, αποφεύγοντας φαινόμενα όπως υπερβολική εκτέλεση άλματος ή πρόωρη αποτυχία.

ΚΕΦΑΛΑΙΟ 5: ΜΕΘΟΔΟΛΟΓΙΑ ΑΞΙΟΛΟΓΗΣΗΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Η παρούσα ενότητα εξετάζει τη διαδικασία αξιολόγησης του εκπαιδευμένου πράκτορα, με βάση τα δεδομένα που καταγράφηκαν κατά την εκπαίδευση μέσω του εργαλείου TensorBoard. Η αξιολόγηση πραγματοποιήθηκε με στόχο να διαπιστωθεί η

αποτελεσματικότητα της στρατηγικής που ανέπτυξε ο πράκτορας, καθώς και η σταθερότητα της πολιτικής που έμαθε. Αρχικά παρουσιάζονται οι βασικές μετρικές που χρησιμοποιήθηκαν και ο ρόλος τους στο πλαίσιο της ενισχυτικής μάθησης, ενώ στη συνέχεια αναλύονται τα αποτελέσματα που προέκυψαν και εξάγονται τα τελικά συμπεράσματα.

5.1 Μεθοδολογία Αξιολόγησης: Μετρικές και ο ρόλος τους

Η αποτίμηση της μαθησιακής πορείας του πράκτορα στηρίχθηκε σε συγκεκριμένες μετρικές που καταγράφηκαν αυτόματα κατά τη διάρκεια της εκπαίδευσης. Οι μετρικές αυτές επιτρέπουν την ποσοτική αξιολόγηση της απόδοσης και τη διερεύνηση της ποιότητας της πολιτικής του πράκτορα, χωρίς να απαιτούνται ξεχωριστά πειράματα σε συνθήκες προσομοίωσης.

Οι βασικές μετρικές που χρησιμοποιήθηκαν ήταν οι εξής:

- **Cumulative Reward (Συσσωρευμένη Ανταμοιβή):** Απεικονίζει τη συνολική ανταμοιβή που συγκεντρώνεται ανά επεισόδιο. Η ανοδική πορεία αυτής της καμπύλης αποτελεί ένδειξη βελτίωσης της πολιτικής και αποτελεσματικής προσαρμογής στο περιβάλλον.
- **Episode Length (Διάρκεια Επεισοδίου):** Μετρά τον αριθμό βημάτων ανά επεισόδιο. Η αυξανόμενη διάρκειά τους συνδέεται με την ικανότητα του πράκτορα να αποφεύγει εμπόδια για μεγαλύτερο χρονικό διάστημα.
- **Policy Loss:** Εκφράζει το σφάλμα προσαρμογής της πολιτικής. Η μείωσή του υποδηλώνει σύγκλιση σε πιο σταθερές στρατηγικές λήψης απόφασης.
- **Value Loss:** Αντιστοιχεί στο σφάλμα εκτίμησης της συνάρτησης αξίας. Η σταδιακή μείωσή του δείχνει ότι ο πράκτορας μαθαίνει να προβλέπει σωστά τη μελλοντική του απόδοση.

Η παρακολούθηση των παραπάνω μετρικών διευκόλυνε τον εντοπισμό μη επιθυμητών συμπεριφορών και συνέβαλε στον εντοπισμό κατάλληλων παραμέτρων εκπαίδευσης, οδηγώντας στην επιλογή του τελικού επιτυχημένου μοντέλου.

5.2 Εκτέλεση Πειραμάτων και Παραγωγή Αποτελεσμάτων

Η διαδικασία εκπαίδευσης του πράκτορα περιλάμβανε αρχικά μια φάση διερεύνησης, κατά την οποία δοκιμάστηκαν διαφορετικοί συνδυασμοί υπερπαραμέτρων του αλγορίθμου PPO. Πολλές από αυτές τις παραμετροποιήσεις οδήγησαν σε μη ικανοποιητικά αποτελέσματα, όπως ασταθή συμπεριφορά ή αδυναμία του πράκτορα να μάθει αποτελεσματική στρατηγική αποφυγής εμποδίων. Μέσα από επαναληπτικές δοκιμές και ανάλυση των μετρικών στο

TensorBoard, προσδιορίστηκε ένα σύνολο ρυθμίσεων που οδήγησε σε σταθερή και σταδιακά βελτιωμένη απόδοση.

Η εκπαίδευση πραγματοποιήθηκε με τη χρήση της λειτουργίας PPO εντός του Unity Editor, ώστε να υπάρχει άμεση παρατήρηση της συμπεριφοράς του πράκτορα κατά τη διάρκεια των επεισοδίων. Το μοντέλο που επιλέχθηκε για αξιολόγηση και χρήση είναι το καλύτερο checkpoint που προέκυψε αυτόματα κατά τη διάρκεια της εκπαίδευσης. Η επιλογή αυτή βασίστηκε στην τιμή της συσσωρευμένης ανταμοιβής και στην ορατή βελτίωση της συμπεριφοράς του πράκτορα εντός του περιβάλλοντος.

Αξιοσημείωτο είναι ότι κατά τη διάρκεια της εκπαίδευσης παρατηρήθηκε ένα φαινόμενο αποδόμησης της πολιτικής (policy degradation) όταν η διαδικασία συνεχίζονταν πέρα από ένα συγκεκριμένο αριθμό βημάτων. Συγκεκριμένα, μετά από ένα σημείο, η τιμή της μέσης ανταμοιβής κατέρρευε απότομα, γεγονός που υποδηλώνει πιθανή υπερεκπαίδευση ή απώλεια σταθερότητας στη μάθηση. Για τον λόγο αυτό, επιλέχθηκε checkpoint που προηγήθηκε αυτής της φάσης, το οποίο παρουσίαζε ισορροπία μεταξύ ανταμοιβής, διάρκειας επεισοδίων και συνέπειας στη συμπεριφορά.

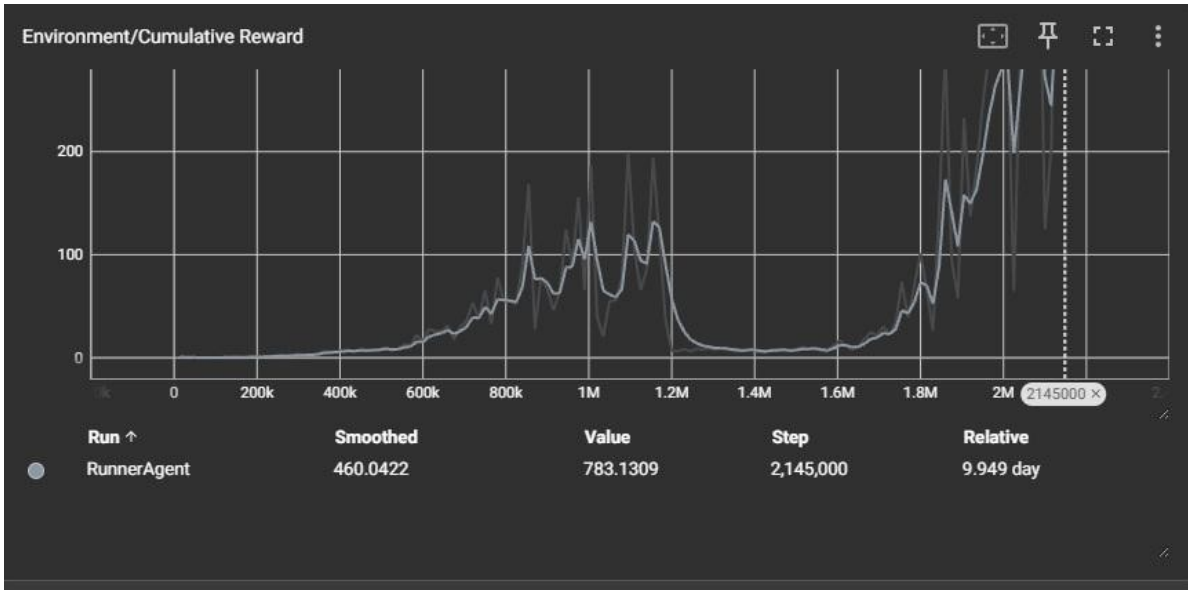
Το τελικό μοντέλο εμφανίζει ικανοποιητική και συνεπή απόδοση, με υψηλά σκορ και σταθερή ικανότητα αποφυγής εμποδίων, επιβεβαιώνοντας την αποτελεσματικότητα της επιλεγμένης διαμόρφωσης παραμέτρων και του σχεδιασμού του συστήματος.

5.3 Ανάλυση και Εξαγωγή Συμπερασμάτων Αξιολόγησης

Η ανάλυση των μετρικών που καταγράφηκαν κατά την εκπαίδευση του πράκτορα προσφέρει ουσιαστική πληροφόρηση για τη σταδιακή εξέλιξη της πολιτικής του και την τελική του απόδοση. Οι σχετικές απεικονίσεις στα Σχήματα 5.1 έως 5.4 αποκαλύπτουν τη δυναμική της μάθησης, τις φάσεις σταθερότητας και αστάθειας, καθώς και τη συνολική ποιότητα του εκπαιδευμένου μοντέλου.

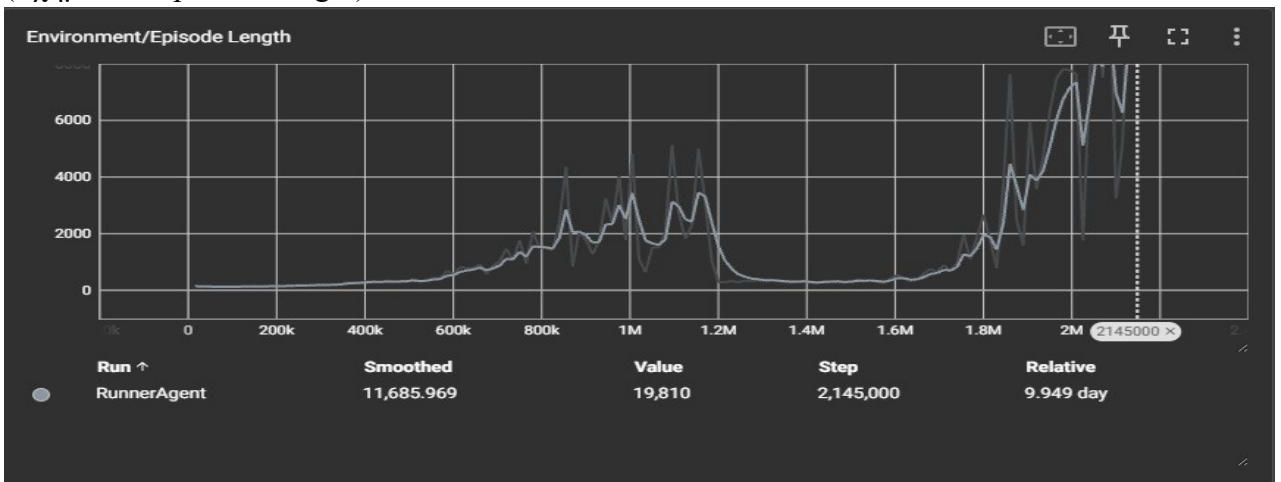
Το Cumulative Reward (Σχήμα 5.1) αποτυπώνει τη συσσωρευμένη ανταμοιβή ανά επεισόδιο και λειτουργεί ως γενικός δείκτης απόδοσης. Αν και παρατηρείται συνολικά ανοδική πορεία, εντοπίζονται περίοδοι σημαντικής υποχώρησης, ιδίως μετά το πρώτο εκατομμύριο βημάτων. Αυτές οι διακυμάνσεις πιθανόν αντανακλούν μεταβατικές φάσεις της πολιτικής, ενδεχομένως λόγω υπερβολικής εξερεύνησης ή προσωρινής αποσταθεροποίησης της εκπαίδευσης. Η σταδιακή αποκατάσταση και η μεταγενέστερη σταθεροποίηση σε υψηλές τιμές ενισχύουν την υπόθεση επιτυχούς σύγκλισης.

(Σχήμα 5.1, Cumulative Reward)



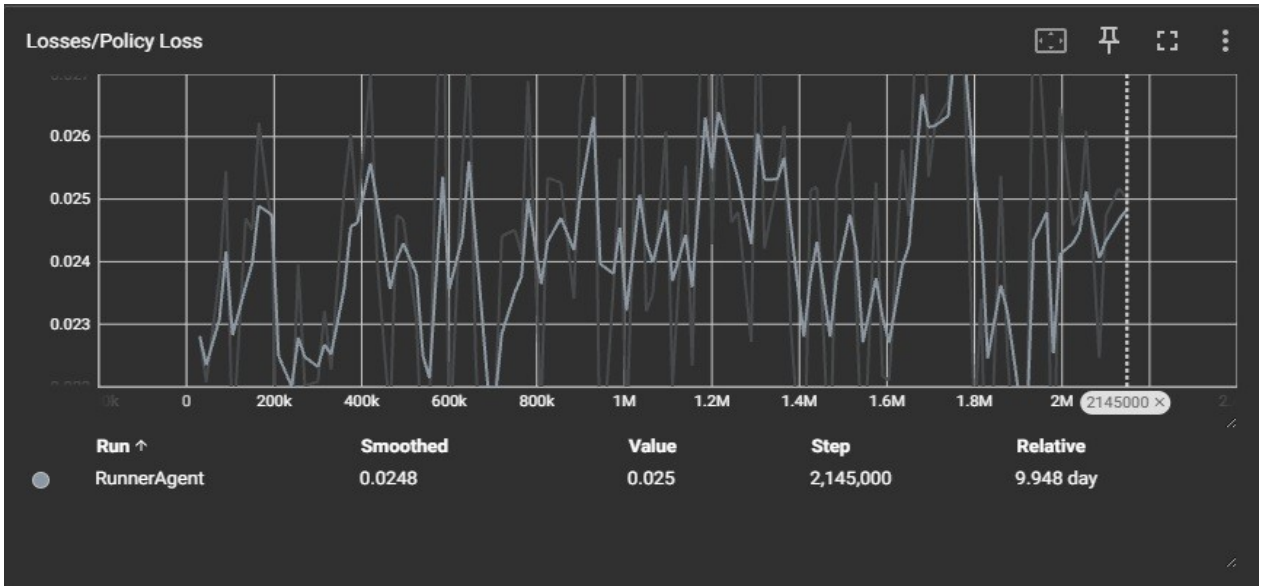
Παρόμοια συμπεριφορά παρατηρείται και στη μέτρηση Episode Length (Σχήμα 5.2), η οποία υποδηλώνει την ικανότητα του πράκτορα να επιβιώνει για περισσότερα βήματα εντός του περιβάλλοντος. Η συσχέτιση της διάρκειας των επεισοδίων με τη συσσωρευμένη ανταμοιβή είναι αναμενόμενη, καθώς η επιτυχής αποφυγή εμποδίων οδηγεί τόσο σε αυξημένο σκορ όσο και σε παράταση του επεισοδίου. Η παρουσία απότομων πτώσεων ακολουθούμενων από φάσεις ανάκαμψης είναι ενδεικτική του συνεχούς επαναπροσδιορισμού της πολιτικής.

(Σχήμα 5.2 Episode Length)



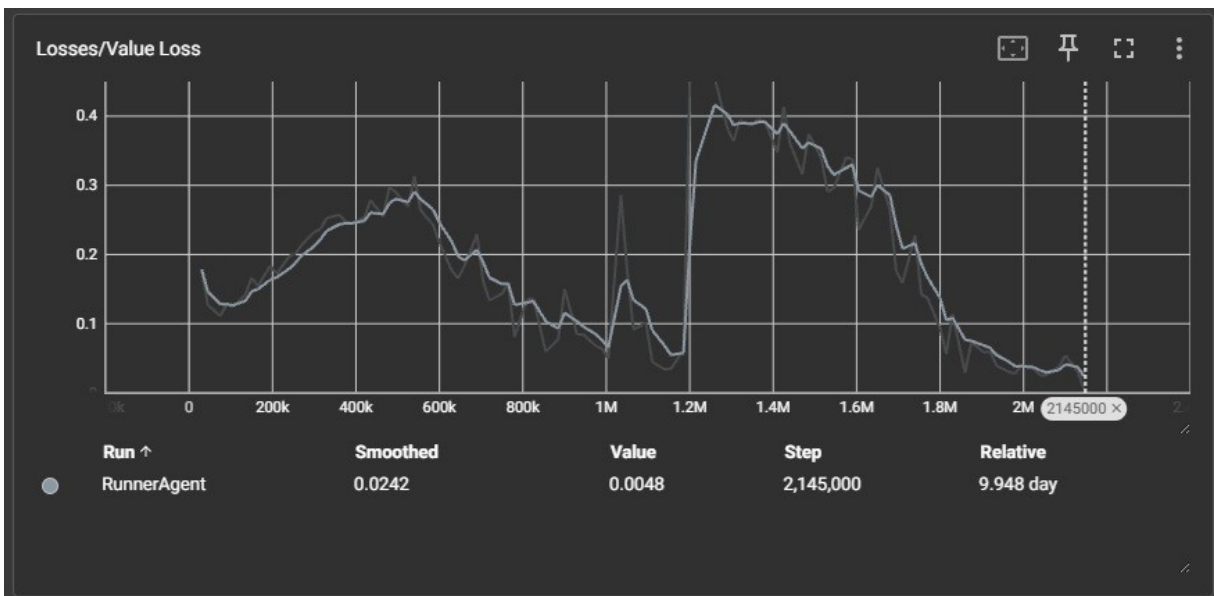
Το Policy Loss (Σχήμα 5.3) παρουσιάζει διακυμάνσεις γύρω από μία σχετικά σταθερή μέση τιμή καθ' όλη τη διάρκεια της εκπαίδευσης. Η συμπεριφορά αυτή συνάδει με τον σχεδιασμό του αλγορίθμου PPO, ο οποίος επιβάλλει περιορισμούς στην αλλαγή της πολιτικής, προκειμένου να διατηρηθεί η σταθερότητα κατά την ενημέρωση των παραμέτρων. Η απουσία ακραίων αυξομειώσεων αποτελεί ένδειξη ότι οι τροποποιήσεις της πολιτικής γίνονται εντός εύλογων ορίων και χωρίς καταστροφικές συνέπειες για τη μαθησιακή διαδικασία.

(Σχήμα 5.3, Policy Loss)



Αντίστοιχα, το Value Loss (Σχήμα 5.4) εμφανίζει σημαντική πτώση μετά από αρχική κορύφωση, γεγονός που αποδεικνύει τη βελτίωση της ικανότητας του δικτύου να εκτιμά την αναμενόμενη ανταμοιβή. Η μετάβαση από ασταθείς προβλέψεις σε χαμηλό και σταθερό σφάλμα αξίας αποτελεί ουσιώδες κριτήριο για την αξιοπιστία των αποφάσεων του πράκτορα, καθώς ένα ακριβές value function ενισχύει τη συνέπεια των πολιτικών επιλογών.

(Σχήμα 5.4, Value Loss)



Αξιοσημείωτο είναι το γεγονός ότι η εκπαίδευση δεν ολοκληρώθηκε βάσει του μέγιστου αριθμού βημάτων, αλλά διακόπηκε νωρίτερα, αφού παρατηρήθηκε φαινόμενο απότομης πτώσης της απόδοσης σε μετέπειτα στάδια. Η συμπεριφορά αυτή ενδέχεται να σχετίζεται με φαινόμενα καταστροφικής λήθης ή αποσταθεροποίησης της πολιτικής (policy degradation), φαινόμενα που αναφέρονται συχνά στη βιβλιογραφία σε περιπτώσεις μακρόχρονης

εκπαίδευσης χωρίς επαρκή κανονικοποίηση. Για τον λόγο αυτό, επελέγη ως τελικό μοντέλο το καλύτερο checkpoint με βάση τη συσσωρευμένη απόδοση.

Συνοψίζοντας, η ανάλυση των διαγραμμάτων καταδεικνύει ότι το σύστημα κατόρθωσε να εκπαιδεύσει έναν πράκτορα με σταθερή και αποτελεσματική στρατηγική. Παρά τις ενδιάμεσες διακυμάνσεις, το μοντέλο συγκλίνει προς αποδοτική συμπεριφορά, επιτυγχάνοντας υψηλές τιμές επιβράβευσης και μακρά διάρκεια επιβίωσης στο περιβάλλον αξιολόγησης.

ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

6.1 Συνολική Αποτίμηση της Εργασίας

Η παρούσα πτυχιακή εργασία ανέδειξε τον ρόλο της Ενισχυτικής Μάθησης σε εφαρμογές παιγνίων, μέσω της ανάπτυξης ενός περιβάλλοντος τύπου endless runner στο Unity και της εκπαίδευσης ενός πράκτορα με τον αλγόριθμο Proximal Policy Optimization (PPO). Η μελέτη συνδύασε θεωρητική θεμελίωση και πρακτική υλοποίηση, παρέχοντας ένα ολοκληρωμένο πλαίσιο που περιλαμβάνει την ανάπτυξη της αρχιτεκτονικής, τον σχεδιασμό του reward system, την παραμετροποίηση του PPO και την αξιολόγηση των αποτελεσμάτων.

Συνολικά, η εργασία κατέδειξε ότι ακόμα και σε περιβάλλοντα με σχετικά απλό χώρο ενεργειών (ένα μόνο άλμα), η ενισχυτική μάθηση είναι ικανή να οδηγήσει σε συμπεριφορές που μιμούνται στρατηγικές ανθρώπινου επιπέδου, με σταδιακή βελτίωση και προσαρμογή μέσω της εμπειρίας.

6.2 Συμπεράσματα από την Αξιολόγηση του Συστήματος

Η ανάλυση των μετρικών εκπαίδευσης (cumulative reward, episode length, policy loss, value loss) ανέδειξε την ικανότητα του πράκτορα να:

- βελτιώνει προοδευτικά την απόδοσή του
- σταθεροποιεί τη στρατηγική του σε επαναλαμβανόμενα επεισόδια
- μαθαίνει να αποφεύγει τυχαία άλματα και να εστιάζει σε αποφάσεις με υψηλή αναμενόμενη ανταμοιβή.

Παρά τις ενδιάμεσες διακυμάνσεις και την εμφάνιση φαινομένων policy degradation σε μακροχρόνια εκπαίδευση, το τελικό μοντέλο παρουσίασε συνεπή και αξιόπιστη απόδοση, επιβεβαιώνοντας τη λειτουργικότητα του αλγορίθμου PPO σε περιβάλλοντα παιγνίων με περιορισμένο αλλά κρίσιμο χώρο αποφάσεων.

6.3 Περιορισμοί και Προβληματισμοί

Η εργασία, αν και επιτυχημένη στην εφαρμογή της, παρουσίασε ορισμένους περιορισμούς:

- Ο πράκτορας διαθέτει περιορισμένο σύνολο ενεργειών (ένα άλμα), γεγονός που απλοποιεί υπερβολικά το πρόβλημα σε σχέση με πιο σύνθετα παιχνίδια
- Ο μηχανισμός ανταμοιβής βασίστηκε σε απλές, γραμμικές τιμές, χωρίς προσαρμογή σε δυναμικές συνθήκες.

- Το περιβάλλον είναι στατικό ως προς τη δυσκολία: η ταχύτητα και η πολυπλοκότητα των εμποδίων παραμένουν σχετικά προβλέψιμες.
- Η διαδικασία εκπαίδευσης ήταν χρονοβόρα, απαιτώντας εκατομμύρια βήματα για τη σταθερή εκμάθηση στρατηγικής.
- Παρατηρήθηκαν φαινόμενα υπερεκπαίδευσης, τα οποία περιόρισαν την αξιοπιστία του μοντέλου σε πολύ μεγάλες περιόδους μάθησης.

Οι παραπάνω περιορισμοί αναδεικνύουν την ανάγκη για περαιτέρω βελτιώσεις σε επίπεδο αλγορίθμων, περιβάλλοντος και μεθοδολογίας αξιολόγησης.

6.4 Προτάσεις για Μελλοντική Βελτίωση και Επεκτάσεις

Με βάση τα συμπεράσματα και τους περιορισμούς, προτείνονται οι ακόλουθες κατευθύνσεις για μελλοντική εργασία:

- **Αύξηση της πολυπλοκότητας του περιβάλλοντος:** ενσωμάτωση περισσότερων ενεργειών (π.χ. διπλό άλμα, ολίσθηση, αλλαγή πορείας), μεταβλητής ταχύτητας εμποδίων και δυναμικών σεναρίων.
- **Βελτιστοποίηση του reward system:** χρήση προσαρμοστικών μηχανισμών ανταμοιβής που λαμβάνουν υπόψη τον βαθμό δυσκολίας της απόφασης ή την ποιότητα του χρονισμού.
- **Δοκιμή εναλλακτικών αλγορίθμων RL:** αξιολόγηση μεθόδων όπως Deep Q-Networks (DQN), Soft Actor-Critic (SAC) ή πιο πρόσφατες προσεγγίσεις μεταμάθησης.
- **Μεταφορά γνώσης (transfer learning):** εκπαίδευση του πράκτορα σε πολλαπλά περιβάλλοντα, με στόχο τη γενίκευση της στρατηγικής.
- **Διαχείριση χρόνου εκπαίδευσης:** αξιοποίηση τεχνικών parallel training ή βελτιστοποιημένων αρχιτεκτονικών νευρωνικών δικτύων για μείωση του χρόνου σύγκλισης.
- **Διερεύνηση πραγματικών εφαρμογών:** μεταφορά της μεθοδολογίας σε τομείς όπως η ρομποτική, η αυτόνομη πλοήγηση και τα adaptive user interfaces, όπου η ανάγκη για λήψη αποφάσεων σε πραγματικό χρόνο είναι κρίσιμη.

Οι παραπάνω προτάσεις ανοίγουν τον δρόμο για μια πιο εκτενή και ρεαλιστική μελέτη της ενισχυτικής μάθησης, επιτρέποντας την περαιτέρω ανάπτυξη πρακτόρων με αυξημένη ευφυΐα, προσαρμοστικότητα και γενίκευση.

BIBΛΙΟΓΡΑΦΙΑ

Kanthraj, S., & Student, B. (2016). Unsupervised Feature Learning and Deep Learning in Artificial Intelligence for Enhancing Cyber Security.

[https://www.idosi.org/mejsr/mejsr24\(10\)16/42.pdf](https://www.idosi.org/mejsr/mejsr24(10)16/42.pdf)

Hasan et al.(2023) Exploring Naive Bayes for Movie Review Sentiment Classification

<https://ijirce.com/admin/main/storage/app/pdf/CR2ZgaqgNkBnmzvEhwRDKK6dWfRHyLAfeaogxNf6.pdf>

Shaveta,. (2023). A review on machine learning. International Journal of Science and Research Archive. 9. 281-285.

https://www.researchgate.net/publication/371258825_A_review_on_machine_learning

Mckinsey & Company. (2024). *What is AI (artificial intelligence)?*

<https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-ai>

Moez Ali(2022) Supervised Machine Learning <https://www.datacamp.com/blog/supervised-machine-learning>

Moranodín-Ahuerma, 2022.<https://ijrpr.com/uploads/V3ISSUE12/IJRPR8827.pdf>

Pakhale, D. V., & Athawale, S. V. (2024) Reinforcement Learning in Machine Learning

<https://www.ijirid.in/3-5-24Oct/3-5-34-Devyani%20Pakhale-Prof%20Samata%20Athawale.pdf>

Zoumana Keita(2024) Classification in Machine Learning: An Introduction.

<https://www.datacamp.com/blog/classification-machine-learning>

A. Lambora, K. Gupta and K. Chopra, "Genetic Algorithm- A Literature Review," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 380-384, doi: [10.1109/COMITCon.2019.8862255](https://doi.org/10.1109/COMITCon.2019.8862255).

Ackermann, M.R., Blömer, J., Kuntze, D. et al. Analysis of Agglomerative Clustering. *Algorithmica* 69, 184–215 (2014). <https://doi.org/10.1007/s00453-012-9717-4>

Aher, S.B., Lobo, L.M.R.J. (2013). Prediction of Course Selection in E-Learning System Using Combined Approach of Unsupervised Learning Algorithm and Association Rule. In: Das, V.V., Chaba, Y. (eds) *Mobile Communication and Power Engineering*. AIM 2012. Communications in Computer and Information Science, vol 296. Springer, Berlin, Heidelberg.https://doi.org/10.1007/978-3-642-35864-7_22

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, 9(8), 1295.

<https://doi.org/10.3390/electronics9081295>

Andrew G. Barto, Reinforcement Learning, IFAC Proceedings Volumes, Volume 31, Issue 29, Supplement 1, 1998, Page 5, ISSN 1474-6670, [https://doi.org/10.1016/S1474-6670\(17\)38315-5](https://doi.org/10.1016/S1474-6670(17)38315-5)

Andrew Y. Ng and Michael I. Jordan. 2001. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01). MIT Press, Cambridge, MA, USA, 841–848.
<https://dl.acm.org/doi/10.5555/2980539.2980648>

Aristawidya, Rafika & Indahwati, Indahwati & Erfiani, Erfiani & Fitrianto, Anwar & Aliu, Muftih. (2024). PERBANDINGAN ANALISIS REGRESI LOGISTIK BINER DAN NAÏVE BAYES CLASSIFIER UNTUK MEMREDIKSI FAKTOR RESIKO DIABETES. Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika. 5. 782-794.
<http://dx.doi.org/10.46306/lb.v5i2.617>

Ashley, K. (2020). Reinforcement Learning in Sports. , 199-219. https://doi.org/10.1007/978-1-4842-5772-2_10.

Awange, J., Paláncz, B., Völgyesi, L. (2020). Neural Networks. In: Hybrid Imaging and Visualization. Springer, Cham. https://doi.org/10.1007/978-3-030-26153-5_5

Bao Chong. K-means clustering algorithm: a brief review. Academic Journal of Computing & Information Science (2021), Vol. 4, Issue 5: 37-40.
<https://doi.org/10.25236/AJCIS.2021.040506>.

Barto, A., & Sutton, R. (1997). Reinforcement Learning in Artificial Intelligence. Advances in psychology, 121, 358-386. [https://doi.org/10.1016/S0166-4115\(97\)80105-7](https://doi.org/10.1016/S0166-4115(97)80105-7).

Bastos, J. A. (2022). Predicting Credit Scores with Boosted Decision Trees. Forecasting, 4(4), 925-935. <https://doi.org/10.3390/forecast4040050>

Becker, Suzanna. (1991). Unsupervised Learning Procedures for Neural Networks. Int. J. Neural Syst.. 2. 17-33. 10.1142/S0129065791000030.
<http://dx.doi.org/10.1142/S0129065791000030>

Bhowmik, Tapan. (2015). Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. Inteligencia Artificial. 18. 14. 10.4114/intartif.vol18iss56pp14-30.
<http://dx.doi.org/10.4114/intartif.vol18iss56pp14-30>

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).
<https://doi.org/10.1023/A:1010933404324>

Brynjolfsson, E., McAfee, A., The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, 9780393239355, 2014 W., W. Norton,
<https://books.google.com/books?id=WiKwAgAAQBAJ>

- Chakraverty, S., Sahoo, D.M., Mahato, N.R. (2019). Hebbian Learning Rule. In: Concepts of Soft Computing. Springer, Singapore. https://doi.org/10.1007/978-981-13-7430-2_12
- Chapelle, Olivier & Schölkopf, Bernhard & Zien, Alexander. (2006). Semi-Supervised Learning. <http://dx.doi.org/10.7551/mitpress/9780262033589.001.0001>.
- Charles, D., Fyfe, C., Livingstone, D., & McGlinchey, S. (2008). Unsupervised Learning in Artificial Neural Networks. , 48-90. <https://doi.org/10.4018/978-1-59140-646-4.CH005>.
- Cialfi, D. (2020). The Self-Organizing Map: An Methodological Note. In: Bucciarelli, E., Chen, SH., Corchado, J. (eds) Decision Economics: Complexity of Decisions and Decisions for Complexity. DECON 2019. Advances in Intelligent Systems and Computing, vol 1009. Springer, Cham. https://doi.org/10.1007/978-3-030-38227-8_28
- Cichosz, P. (1994). Truncating temporal differences: On the efficient implementation of TD (λ) for reinforcement learning. Journal of Artificial Intelligence Research, 2, 287-318. <https://doi.org/10.1613/jair.135>
- Cios, K.J., Swiniarski, R.W., Pedrycz, W., Kurgan, L.A. (2007). Unsupervised Learning: Association Rules. In: Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-36795-8_10
- Co-Reyes, J. D., Miao, Y., Peng, D., Real, E., Levine, S., Le, Q. V., ... & Faust, A. (2021). Evolving reinforcement learning algorithms. arXiv preprint [arXiv:2101.03958](https://arxiv.org/abs/2101.03958).
- Coadou, Y. (2022). Boosted decision trees. In Artificial Intelligence for High Energy Physics (pp. 9-58). <https://doi.org/10.48550/arXiv.2206.09645>
- Cragg, J.G. (1990). Monte Carlo Methods. In: Eatwell, J., Milgate, M., Newman, P. (eds) Time Series and Statistics. The New Palgrave. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-20865-4_22
- D. Sánchez, M.A. Vila, L. Cerda, J.M. Serrano, Association rules applied to credit card fraud detection, Expert Systems with Applications, Volume 36, Issue 2, Part 2, 2009, Pages 3630-3640, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2008.02.001>.
- Dai, J., Ji, J., Yang, L., Zheng, Q., & Pan, G. (2023). Augmented Proximal Policy Optimization for Safe Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 37(6), 7288-7295. <https://doi.org/10.1609/aaai.v37i6.25888>
- Di Felice, M., Deroche, A., Trupkin, I., Chatterjee, P., & Cattaneo, M. (2023). Depression and Anxiety Diagnosis Using Unsupervised Learning Approach. , 12-24. https://ceur-ws.org/Vol-3520/icaiw_waai_2.pdf
- Diederichs, E. (2019). Reinforcement Learning - A Technical Introduction. Journal of Autonomous Intelligence. <https://doi.org/10.32629/JAI.V2I2.45>.

- Diederichs, E. (2019). Reinforcement Learning - A Technical Introduction. Journal of Autonomous Intelligence. <https://doi.org/10.32629/JAI.V2I2.45>.
- Duryea, E. , Ganger, M. and Hu, W. (2016) Exploring Deep Reinforcement Learning with Multi Q-Learning. Intelligent Control and Automation, 7, 129-144. doi: <http://dx.doi.org/10.4236/ica.2016.74012>.
- Even-Dar, E. (2016). Reinforcement Learning. In: Kao, MY. (eds) Encyclopedia of Algorithms. Springer, New York, NY. https://doi.org/10.1007/978-1-4939-2864-4_341
- F. AlMahamid and K. Grolinger, "Reinforcement Learning Algorithms: An Overview and Classification," 2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), ON, Canada, 2021, pp. 1-7, <https://doi.org/10.1109/CCECE53047.2021.9569056>.
- Fan, J., Wang, Z., Xie, Y. & Yang, Z.. (2020). A Theoretical Analysis of Deep Q-Learning. Proceedings of the 2nd Conference on Learning for Dynamics and Control, in Proceedings of Machine Learning Research 120:486-489 Available from <https://proceedings.mlr.press/v120/yang20a.html>.
- Ghasemi, M., & Ebrahimi, D. (2024). Introduction to Reinforcement Learning. <https://arxiv.org/pdf/2408.07712>
- Ghasemi, Majid & Moosavi, Amir & Sorkhoh, Ibrahim & Agrawal, Anjali & Alzhouri, Fadi & Ebrahimi, Dariush. (2024). An Introduction to Reinforcement Learning: Fundamental Concepts and Practical Applications. <http://dx.doi.org/10.48550/arXiv.2408.07712>
- Gu, Yang & Chen, C. & Wang, Xuesong. (2021). Proximal Policy Optimization With Policy Feedback. IEEE Transactions on Systems, Man, and Cybernetics: Systems. PP. 1-11. 10.1109/TSMC.2021.3098451. https://www.researchgate.net/publication/353563772_Proximal_Policy_Optimization_With_Policy_Feedback
- Herbin, M., Bonnet, N. (2002). An Improved Method for Estimating the Modes of the Probability Density Function and the Number of Classes for PDF-based Clustering. In: Jajuga, K., Sokołowski, A., Bock, HH. (eds) Classification, Clustering, and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-56181-8_28
- Herrera, F. et al. (2016). Unsupervised Multiple Instance Learning. In: Multiple Instance Learning. Springer, Cham. https://doi.org/10.1007/978-3-319-47759-6_7
- I. Grondman, L. Busoniu, G. A. D. Lopes and R. Babuska, "A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 6, pp. 1291-1307, Nov. 2012, doi: [10.1109/TSMCC.2012.2218595](https://doi.org/10.1109/TSMCC.2012.2218595).

- Jia, Y., & Zhou, X. Y. (2022). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275), 1-50.
<https://doi.org/10.48550/arXiv.2111.11232>
- Jijo, Bahzad & Abdulazeez, Adnan. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*. 2. 20-28.
https://www.researchgate.net/publication/350386944_Classification_Based_on_Decision_Tree_Algorithm_for_Machine_Learning
- John McCarthy(2007) What is Artificial Intelligence?
https://www.researchgate.net/publication/28762490_What_is_Artificial_Intelligence
- Joshi, Srivatsa & Kumar, Vishwas & Venkataramanan, Vishaka & C S, Kaliprasad. (2023). A Review on Neural Networks and its Applications. *Journal of Computer Technology & Applications*. 14. 2023. 10.37591/jocta.v14i2.1062.
<http://dx.doi.org/10.37591/jocta.v14i2.1062>
- Jung, A. (2018). A Gentle Introduction to Supervised Machine Learning. [ArXiv, abs/1805.05052](https://arxiv.org/abs/1805.05052).
- K. . -R. Hsieh and W. . -T. Chen, "A neural network model which combines unsupervised and supervised learning," in *IEEE Transactions on Neural Networks*, vol. 4, no. 2, pp. 357-360, March 1993, doi: 10.1109/72.207624. <https://doi.org/10.1109/72.207624>
- Kaelbling, L., Littman, M., & Moore, A. (1995). An Introduction to Reinforcement Learning. , 90-127. https://doi.org/10.1007/978-3-642-79629-6_5.
- Kim Larsen. 2005. Generalized Naive Bayes Classifiers. *SIGKDD Explor. Newsl.* 7, 1 (June 2005), 76–81. <https://doi.org/10.1145/1089815.1089826>
- Kote, V. (2019). Unsupervised-Learning Assisted Artificial Neural Network for Optimization. . <https://doi.org/10.25777/KHDW-4A23>.
- Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., Steinbrecher, M. (2016). Introduction to Neural Networks. In: *Computational Intelligence. Texts in Computer Science*. Springer, London. https://doi.org/10.1007/978-1-4471-7296-3_2
- Kunz, F. (2000). An introduction to temporal difference learning. In *Seminar on autonomous learning systems* (pp. 21-22).
[https://www.ice.ci.ritsumei.ac.jp/~ruck/CLASSES/INTELISYS/An_Introduction_to_Temporal_Difference_Learning_\(Kunz_ALS_2013_revised_by_RT\).pdf](https://www.ice.ci.ritsumei.ac.jp/~ruck/CLASSES/INTELISYS/An_Introduction_to_Temporal_Difference_Learning_(Kunz_ALS_2013_revised_by_RT).pdf)
- Kurth-Nelson Z, Redish AD (2009) Temporal-Difference Reinforcement Learning with Distributed Representations. *PLoS ONE* 4(10): e7362.
<https://doi.org/10.1371/journal.pone.0007362>

- Lee, Kidong & Booth, David & Alam, Pervaiz. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications*. 29. 1-16. 10.1016/j.eswa.2005.01.004.
<http://dx.doi.org/10.1016/j.eswa.2005.01.004>
- Leela, S. (2011). Enhancing K-Means Clustering Algorithm.
https://www.semanticscholar.org/paper/Enhancing-K-Means-Clustering-Algorithm-Leela/Od4274b71ad5ae9ec041e4b8dab06cc6deb5591a?utm_source=direct_link
- Lote, S., B, P., & Patrer, D. (2020). Neural networks for machine learning applications. *World Journal of Advanced Research and Reviews*. <https://doi.org/10.30574/wjarr.2020.6.1.0055>.
- Ma, J. (2024, March). Discerning temporal difference learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 13, pp. 14238-14245).
<https://doi.org/10.48550/arXiv.2310.08091>
- Maña, C. (2017). Monte Carlo Methods. In: *Probability and Statistics for Particle Physics. UNITEXT for Physics*. Springer, Cham. https://doi.org/10.1007/978-3-319-55738-0_3
- Mascagni, Michael & Simonov, Nikolai. (2004). Monte Carlo Methods for Calculating Some Physical Properties of Large Molecules. *SIAM J. Scientific Computing*. 26. 339-357. 10.1137/S1064827503422221. <http://dx.doi.org/10.1137/S1064827503422221>
- Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. *Nature* 518, 529–533 (2015). <https://doi.org/10.1038/nature14236>
- N. Ponomareva, T. Colthurst, G. Hendry, S. Haykal and S. Radpour, "Compact multi-class boosted trees," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 47-56, <https://doi.org/10.1109/BigData.2017.8257910>.
- Nkemdilim, M., Uzoamaka, P., Daniel, U., & Chidi, M. (2024). An Overview of Supervised Machine Learning Paradigms and their Classifiers. *International Journal of Advanced Engineering, Management and Science*.<https://doi.org/10.22161/ijaems.103.4>.
- Oh, J., Hessel, M., Czarnecki, W. M., Xu, Z., van Hasselt, H. P., Singh, S., & Silver, D. (2020). Discovering reinforcement learning algorithms. *Advances in Neural Information Processing Systems*, 33, 1060-1070 <https://arxiv.org/pdf/2007.08794v1>
- Pandey, P., Pandey, D., & Kumar, S. (2010). Reinforcement learning by comparing immediate reward. *arXiv preprint* <https://doi.org/10.48550/arXiv.1009.2566>.
- Pironneau, O. (2021). Supervised Learning and Applied Mathematics. *Intelligent Systems, Control and Automation: Science and Engineering*. https://doi.org/10.1007/978-3-030-70787-3_4.

- Pradhan, Shreeja. (2024). Evaluating Deep Reinforcement Learning Algorithms. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. 08. 1-6. 10.55041/IJSREM37434.
<http://dx.doi.org/10.55041/IJSREM37434>
- Quinto, B. (2020). Unsupervised Learning. In: Next-Generation Machine Learning with Spark. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-5669-5_4
- Ran, X., Xi, Y., Lu, Y. et al. Comprehensive survey on hierarchical clustering algorithms and the recent developments. Artif Intell Rev 56, 8219–8264 (2023).
<https://doi.org/10.1007/s10462-022-10366-3>
- Ray H. White, Competitive hebbian learning: Algorithm and demonstrations, Neural Networks, Volume 5, Issue 2, 1992, Pages 261-275, ISSN 0893-6080.
[https://doi.org/10.1016/S0893-6080\(05\)80024-3](https://doi.org/10.1016/S0893-6080(05)80024-3).
- Reddy, M & Makara, Vivekananda & R U V N, Satish. (2017). Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering. 5.
https://www.researchgate.net/publication/326200086_Divisive_Hierarchical_Clustering_with_K-means_and_Agglomerative_Hierarchical_Clustering
- Rokach, Lior & Maimon, Oded. (2005). Top-Down Induction of Decision Trees Classifiers—A Survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 35. 476 - 487. <http://dx.doi.org/10.1109/TSMCC.2004.843247>.
- Rougier, N. P., & Detorakis, G. I. (2021). Randomized Self-Organizing Map. Neural Computation, 33(8), 2241-2273. <https://doi.org/10.48550/arXiv.2011.09534>
- Russell, S. & Norvig, Peter. (2003). Artificial Intelligence, A Modern Approach. Second Edition.
https://www.researchgate.net/publication/272161464_Artificial_Intelligence_A_Modern_Approach_Second_Edition
- Sakai, T., Komazaki, T., Imiya, A. (2007). Scale-Space Clustering with Recursive Validation. In: Sgallari, F., Murli, A., Paragios, N. (eds) Scale Space and Variational Methods in Computer Vision. SSVM 2007. Lecture Notes in Computer Science, vol 4485. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72823-8_25
- Sandhu, M., Saini, A., Kaur, G., & , P. (2024). REINFORCEMENT LEARNING: FRAMEWORK, APPLICATIONS AND CHALLENGES. International Journal of Engineering Science and Humanities. <https://doi.org/10.62904/s3qf5660>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. <https://arxiv.org/abs/1707.06347>.

- Serra, Angela & Tagliaferri, Roberto. (2018). Unsupervised Learning: Clustering. <http://dx.doi.org/10.1016/B978-0-12-809633-8.20487-1>
- Sharma, R. (2020). Study of Supervised Learning and Unsupervised Learning. *International Journal for Research in Applied Science and Engineering Technology*, 8, 588-593. <https://doi.org/10.22214/ijraset.2020.6095>.
- Sharma, R., Saxena, K., & Rana, A. (2021). Unsupervised Learning in Accordance With New Aspects of Artificial Intelligence. *Machine Learning Approach for Cloud Data Analytics in IoT*. <https://doi.org/10.1002/9781119785873.ch17>.
- Shen, Keyi & Tian, Ye & Hu, Bisong & Luo, Jin & Qi, Shuhua & Chen, Songli & Lin, Hui. (2024). Association rule mining of air quality through an improved Apriori algorithm: A case study in 244 Chinese cities. *Transactions in GIS*. 28. 726-745. 10.1111/tgis.13156. <http://dx.doi.org/10.1111/tgis.13156>
- Shimpi, P. (2025). Reinforcement Learning in Real Life Applications. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*. <https://doi.org/10.55041/ijsrem40881>.
- Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, 23(6), 759. <https://doi.org/10.3390/e23060759>
- Sindhu Meena, K., Suriya, S. (2020). A Survey on Supervised and Unsupervised Learning Techniques. In: Kumar, L., Jayashree, L., Manimegalai, R. (eds) *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications. AISGSC 2019 2019*. Springer, Cham. https://doi.org/10.1007/978-3-030-24051-6_58
- Stephanie Forrest. 1996. Genetic algorithms. *ACM Comput. Surv.* 28, 1 (March 1996), 77–80. <https://doi.org/10.1145/234313.234350>
- Subramanian, A., Chitlangia, S., & Baths, V. (2020). Reinforcement learning and its connections with neuroscience and psychology. *Neural networks : the official journal of the International Neural Network Society*, 145, 271-287 . <https://doi.org/10.1016/j.neunet.2021.10.003>.
- Suprianto Panjaitan et al 2019 *J. Phys.: Conf. Ser.* 1255 012057 <https://iopscience.iop.org/article/10.1088/1742-6596/1255/1/012057>
- Sutton, R., & Barto, A. (2005). Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, 16, 285-286. <https://doi.org/10.1109/TNN.1998.712192>.
- Sutton, R.S. Introduction: The challenge of reinforcement learning. *Mach Learn* 8, 225–227 (1992). <https://doi.org/10.1007/BF00992695>

- Szepesvari, C. (2010). Algorithms for Reinforcement Learning. . <https://doi.org/10.1007/978-3-031-01551-9>.
- Taunk, Kashvi & De, Sanjukta & Verma, Srishti & Swetapadma, Aleena. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 1255-1260. 10.1109/ICCS45141.2019.9065747. <http://dx.doi.org/10.1109/ICCS45141.2019.9065747>
- Taylor, J.G. (2002). Neural Networks. In: Shadbolt, J., Taylor, J.G. (eds) Neural Networks and the Financial Markets. Perspectives in Neural Computing. Springer, London. https://doi.org/10.1007/978-1-4471-0151-2_11
- van Engelen, J.E., Hoos, H.H. A survey on semi-supervised learning. Mach Learn 109, 373–440 (2020). <https://doi.org/10.1007/s10994-019-05855-6>
- Verma, R., Nagar, V., & Mahapatra, S. (2021). Introduction to Supervised Learning. Data Analytics in Bioinformatics. <https://doi.org/10.1002/9781119785620.CH1>.
- Wang, Pei. (2008). What Do You Mean by “AI”?. Frontiers in Artificial Intelligence and Applications. 171. 362-373. https://www.researchgate.net/publication/262357941_What_Do_You_Mean_by_AI
- Wang, Y., He, H. & Tan, X.. (2020). Truly Proximal Policy Optimization. Proceedings of The 35th Uncertainty in Artificial Intelligence Conference, in Proceedings of Machine Learning Research 115:113-122 Available from <https://proceedings.mlr.press/v115/wang20b>
- Wang, Y., He, H., & Tan, X. (2020, August). Truly proximal policy optimization. In Uncertainty in artificial intelligence (pp. 113-122). PMLR. <https://arxiv.org/pdf/1903.07940>
- WILLIAMS, R. J., & PENG, J. (1991). Function Optimization using Connectionist Reinforcement Learning Algorithms. Connection Science, 3(3), 241–268. <https://doi.org/10.1080/09540099108946587>
- Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8, 229–256 (1992). <https://doi.org/10.1007/BF00992696>
- Xujun Peng, Srirangaraj Setlur, Venu Govindaraju, Sitaram Ramachandrupa, Using a boosted tree classifier for text segmentation in hand-annotated documents, Pattern Recognition Letters, Volume 33, Issue 7, 2012, Pages 943-950, ISSN 0167-8655, <https://doi.org/10.1016/j.patrec.2011.09.007>.
- Y. P. Awate, "Policy-Gradient Based Actor-Critic Algorithms," 2009 WRI Global Congress on Intelligent Systems, Xiamen, China, 2009, pp. 505-509, doi: <https://doi.org/10.1109/GCIS.2009.372>.

Yedavalli, V., Tong, E., Martin, D., Yeom, K., & Forkert, N. (2020). Artificial intelligence in stroke imaging: Current and future perspectives.. *Clinical imaging*, 69, 246-254.
<https://doi.org/10.1016/j.clinimag.2020.09.005>.

Yu, Chengcheng & Zhang, Lijun & Yin, Dawei & Peng, Dezhong & Huang, Haixiao. (2021). Proximal Policy Optimization with Future rewards. *Journal of Physics: Conference Series*. 2010. 012085. 10.1088/1742-6596/2010/1/012085. <http://dx.doi.org/10.1088/1742-6596/2010/1/012085>

Yulong Zhang, Li Chen, Xingxing Liang, Jing Yang, Yang Ding, Yanghe Feng, "AlphaStar: an integrated application of reinforcement learning algorithms," Proc. SPIE 12288, International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2022), 1228816 (2 December 2022); <https://doi.org/10.1117/12.2641019>

Zhang, L., Shen, L., Yang, L., Chen, S., Yuan, B., Wang, X., & Tao, D. (2022). Penalized proximal policy optimization for safe reinforcement learning <https://arxiv.org/abs/2205.11814>

Zhu, B., & Shoaran, M. (2021). Tree in tree: from decision trees to decision graphs. *Advances in Neural Information Processing Systems*, 34, 13707-13718.
<https://doi.org/10.48550/arXiv.2110.00392>