



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ

Σύγκριση αλγορίθμων μηχανικής μάθησης για την εκπαίδευση ευφυών πρακτόρων σε περιβάλλον παιχνιδιού

Comparison of machine learning algorithms for the training of intelligent agents in a game environment

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Τσιάλιος
Αριθμός Μητρώου: 1072868

Επιβλέπων
Κυριάκος Σγάρμπας, Αναπληρωτής Καθηγητής

Πάτρα
Σεπτέμβριος 2024

Πανεπιστήμιο Πατρών, Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών.

©2024 —Με την επιφύλαξη παντός δικαιώματος

Το σύνολο της εργασίας αποτελεί πρωτότυπο έργο, παραχθέν από τον Γεώργιο Τσιάλιο, και δεν παραβιάζει δικαιώματα τρίτων καθ' οιονδήποτε τρόπο. Αν η εργασία περιέχει υλικό, το οποίο δεν έχει παραχθεί από τον ίδιο, αυτό είναι ευδιάκριτο και αναφέρεται ρητώς εντός του κειμένου της εργασίας ως προϊόν εργασίας τρίτου, σημειώνοντας με παρομοίως σαφή τρόπο τα στοιχεία ταυτοποίησής του, ενώ παράλληλα βεβαιώνει πως στην περίπτωση χρήσης αυτούσιων γραφικών αναπαραστάσεων, εικόνων, γραφημάτων κ.λπ., έχει λάβει τη χωρίς περιορισμούς άδεια του κατόχου των πνευματικών δικαιωμάτων για την συμπερίληψη και επακόλουθη δημοσίευση του υλικού αυτού.

Γεώργιος Τσιάλιος

ΠΙΣΤΟΠΟΙΗΣΗ

Πιστοποιείται ότι η διπλωματική εργασία με θέμα
**Σύγκριση αλγορίθμων μηχανικής μάθησης για την εκπαίδευση
ευφυών πρακτόρων σε περιβάλλον παιχνιδιού**
του φοιτητή του τμήματος Ηλεκτρολόγων Μηχανικών & Τεχνολογίας
Υπολογιστών

Γεωργίου Τσιάλιου του Ιωάννη

Αριθμός Μητρώου: 1072868

παρουσιάστηκε δημόσια στο τμήμα Ηλεκτρολόγων Μηχανικών &
Τεχνολογίας Υπολογιστών στις

13/9/2024

και εξετάστηκε από την ακόλουθη εξεταστική επιτροπή:

Κυριάκος Σγάρμπας, Αναπληρωτής Καθηγητής, ΤΗΜ&ΤΥ (επιβλέπων)
Σοφία Δασκαλάκη, Επίκουρη Καθηγήτρια, ΤΗΜ&ΤΥ (μέλος επιτροπής)
Χρήστος Φείδας, Αναπληρωτής Καθηγητής, ΤΗΜ&ΤΥ (μέλος επιτροπής)

Ο Επιβλέπων

Ο Διευθυντής του Τομέα

Κυριάκος Σγάρμπας
Αναπληρωτής Καθηγητής

Μιχαήλ Λογοθέτης
Αναπληρωτής Καθηγητής

Σύνοψη

Στη σημερινή εποχή, το επιστημονικό πεδίο της Τεχνητής Νοημοσύνης αποτελεί ένα από τα πιο ραγδαία αναπτυσσόμενα ερευνητικά αντικείμενα παγκοσμίως. Με πιο πρόσφατο παράδειγμα την ανάπτυξη των μεγάλων γλωσσικών μοντέλων (LLMs) όπως το ChatGPT της OpenAI, η Τεχνητή Νοημοσύνη παρεισφρύει ολοένα και περισσότερο στη ζωή των ανθρώπων, παρέχοντας εφαρμογές που λύνουν προβλήματα της καθημερινότητας με υπεράνθρωπη ακρίβεια και ταχύτητα. Στον πυρήνα των εφαρμογών αυτών βρίσκονται συχνά αλγόριθμοι Μηχανικής Μάθησης, ενός υποπεδίου της Τεχνητής Νοημοσύνης. Οι συγκεκριμένοι αλγόριθμοι εξετάζονται συνήθως πρώτα σε δοκιμαστικά περιβάλλοντα, όπως παιχνίδια, όπου η προσομοίωση της πραγματικότητας είναι εύκολη και ακίνδυνη.

Στο πλαίσιο αυτό, στόχος της παρούσας διπλωματικής εργασίας είναι η εκπαίδευση πρακτόρων τεχνητής νοημοσύνης σε ένα απλό παιχνίδι, χρησιμοποιώντας διαφορετικούς αλγορίθμους και η σύγκριση τους σε όρους χρόνου εκπαίδευσης και τελικής επίδοσης. Συγκεκριμένα, αναπτύχθηκε ένα παιχνίδι στο οποίο ο πράκτορας καλείται να παρκάρει ένα αμάξι σε μία τυχαία θέση στάθμευσης. Οι αλγόριθμοι που εξετάστηκαν ανήκουν στην υποκατηγορία της Μηχανικής Μάθησης που ονομάζεται Ενισχυτική Μάθηση και είναι οι εξής: Q-learning, Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), Soft Actor Critic (SAC) και Twin Delayed Deep Deterministic Policy Gradient (TD3).

Ο κώδικας που αναπτύχθηκε για την υλοποίηση του παιχνιδιού και την εκπαίδευση των πρακτόρων είναι ελεύθερα διαθέσιμος στον παρακάτω σύνδεσμο: [GitHub Repository](#).

Λέξεις-κλειδιά: Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Βαθιά Ενισχυτική Μάθηση, Νευρωνικά Δίκτυα, Αυτοδηγούμενα οχήματα, OpenAI Gymnasium, Stable Baselines 3

Abstract

In today's era, the scientific field of Artificial Intelligence is one of the most rapidly evolving research areas worldwide. Artificial Intelligence is increasingly entering people's lives, by offering applications that solve everyday problems with superhuman accuracy and speed. A recent example of this, is the development of Large Language Models (LLMs) like OpenAI's ChatGPT. At the core of these applications are often Machine Learning algorithms, a subfield of Artificial Intelligence. These algorithms are usually tested first in experimental environments, such as games, where simulating reality is easy and safe.

In this context, the goal of this thesis is to train Artificial Intelligence agents in a simple game using different algorithms and compare them in terms of training time and final performance. More specifically, a game was developed in which the agent is tasked with parking a car in a random parking spot. The following algorithms -which belong to the Machine Learning subcategory called Reinforcement Learning- were used: Q-learning, Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), Soft Actor Critic (SAC), and Twin Delayed Deep Deterministic Policy Gradient (TD3).

The code developed for the implementation of the game and the training of the agents is publicly available at the following link: [GitHub Repository](#).

Keywords: Artificial Intelligence, Machine Learning, Deep Reinforcement Learning, Neural Networks, Self-driving cars, OpenAI Gymnasium, Stable Baselines 3

Ευχαριστίες

Φτάνοντας στο τέλος της ακαδημαϊκής μου σταδιοδρομίας, θα ήθελα να ευχαριστήσω όλους όσους με βοήθησαν, έμπρακτα αλλά και όχι μόνο, τα τελευταία πέντε έτη των σπουδών μου.

Αρχικά, επιθυμώ να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας, κο Κυριάκο Σγάρμπα, ο οποίος μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον ερευνητικό πεδίο. Ακόμα, η εμπιστοσύνη που μου έδειξε και η καθοδήγηση του έπαιξαν καθοριστικό ρόλο στην ομαλή εκπόνηση της παρούσας εργασίας.

Στη συνέχεια, θα ήθελα να ευχαριστήσω θερμά την οικογένεια και τους φίλους μου για την συνεχή στήριξη που μου προσφέρουν. Ειδική αναφορά επιθυμώ να κάνω σε δύο άτομα, στα οποία οφείλω, σε μεγάλο βάθμο, τη μέχρι τώρα πορεία μου. Αρχικά, ευχαριστώ τον πατέρα μου, Ιωάννη, για τη μόνψη υποστήριξη και την ανεκτίμητη συμπαράσταση του. Έπειτα, ευχαριστώ τον εξαιρετικό συνάδελφο και φίλο, Χρήστο Κατσανδρή, ο οποίος αποτέλεσε για εμένα παράδειγμα προς μίμηση και με ενέπνευσε να γίνω καλύτερος. Η αδιάλειπτη προθυμία του να βοηθήσει αποδείχθηκε πολύτιμη πολλές φορές, ενώ η συνεργασία μας σε διάφορα μαθήματα και εργασίες ήταν χαρά και τιμή μου.

Περιεχόμενα

| | |
|---|----------|
| 1 Εισαγωγή | 1 |
| 1.1 Στόχος | 1 |
| 1.2 Διάρθρωση Διπλωματικής Εργασίας | 2 |
| 2 Επισκόπηση του χώρου | 3 |
| 2.1 Αυτόνομα οχήματα | 3 |
| 2.1.1 Ορισμός | 3 |
| 2.1.2 Αρχές λειτουργίας | 3 |
| 2.1.3 Πλεονεκτήματα και Μειονεκτήματα | 5 |
| 2.1.4 Η κατάσταση σήμερα | 6 |
| 2.2 Προηγούμενη έρευνα | 7 |
| 3 Βασική Θεωρία | 9 |
| 3.1 Τεχνητή Νοημοσύνη | 9 |
| 3.1.1 Ορισμός | 9 |
| 3.1.2 Μοντέλα και Πράκτορες τεχνητής νοημοσύνης | 9 |
| 3.1.3 Ιστορική Εξέλιξη | 10 |
| 3.1.4 Κατηγορίες | 11 |
| 3.1.5 Εφαρμογές | 14 |
| 3.2 Μηχανική Μάθηση | 15 |
| 3.2.1 Ορισμός | 15 |
| 3.2.2 Κατάταξη πεδίου | 16 |
| 3.2.3 Κατηγορίες | 17 |
| 3.3 Ενισχυτική Μάθηση | 19 |
| 3.3.1 Γενική επισκόπηση | 19 |
| 3.3.2 Βασικές Έννοιες και Ορολογία | 23 |
| 3.3.3 Κατηγορίες αλγορίθμων | 33 |
| 3.3.4 Ο αλγόριθμος Q -learning | 39 |
| 3.4 Βαθιά Ενισχυτική Μάθηση | 47 |
| 3.4.1 Ορισμός και Χαρακτηριστικά | 47 |
| 3.4.2 Τεχνητά Νευρωνικά Δίκτυα | 48 |

| | | |
|-------|-----------------------------|-----------|
| 3.4.3 | Ο αλγόριθμος PPO | 59 |
| 3.4.4 | Ο αλγόριθμος DDPG | 62 |
| 3.4.5 | Ο αλγόριθμος TD3 | 64 |
| 3.4.6 | Ο αλγόριθμος SAC | 67 |
| 3.5 | Σύνοψη | 69 |
| | Βιβλιογραφία | 71 |

Κατάλογος πινάκων

Κατάλογος σχημάτων

| | | |
|---------|--|----|
| 2.1 | Επίπεδα αυτοματισμού αυτόνομων οχημάτων (Society of Automobile Engineers 2021). | 4 |
| 2.2 | Τεχνολογίες αυτόνομων οχημάτων (University of Michigan Center for Sustainable Systems 2023) | 5 |
| 3.1 | Κατηγορίες τεχνητής νοημοσύνης (Lateef 2024). | 12 |
| 3.2 | Ιεραρχία πεδίων τεχνητής νοημοσύνης (Zand κ.ά. 2022). | 16 |
| 3.3 | Κατηγορίες μηχανικής μάθησης (Stewart 2023). | 17 |
| 3.4 | Κύκλος Ενισχυτικής Μάθησης (Lee 2019). | 24 |
| 3.5 | Παράδειγμα Διαδικασίας Απόφασης Μαρκόβ (Alvarez 2017) | 25 |
| 3.6 | Το δίλημμα Εξερεύνησης - Αξιοποίησης (Parkinson 2019). | 31 |
| 3.7 | Ταξινόμηση αλγορίθμων ενισχυτικής μάθησης (OpenAI 2018). | 34 |
| 3.8 | Εκτενέστερη ταξινόμηση αλγορίθμων ενισχυτικής μάθησης (Prijono 2020). | 38 |
| 3.9 | Πίνακας Q για την αποθήκευση των τιμών της συνάρτησης $Q(s, a)$ (Baeldung 2023). . | 40 |
| 3.10 | Η τεχνική ϵ -greedy (Baeldung 2023). | 42 |
| 3.11 | Το παιχνίδι Frozen Lake της βιβλιοθήκης OpenAI Gymnasium. | 44 |
| 3.12 | Αποτελέσματα εκπαίδευσης στο παιχνίδι Frozen Lake (Szymanski 2018). | 45 |
| 3.13 | Κλασική εναντίον Βαθιάς Ενισχυτικής Μάθησης (Quang 2024). | 47 |
| 3.14 | Σύγκριση βιολογικού και τεχνητού νευρώνα (Karpfathy 2016). | 49 |
| 3.15 | Γραφική παράσταση της συνάρτησης Tanh. | 50 |
| 3.16 | Γραφική παράσταση της συνάρτησης ReLU. | 51 |
| 3.17 | Παράδειγμα τεχνητού νευρωνικού δικτύου (Comsol 2023). | 52 |
| 3.18 | Αλληλεπίδραση δικτύων Δράστη-Κριτή (Sutton και Barto 2018). | 59 |
| 3.19 | Συχνότητα χρήσης αλγορίθμων ενισχυτικής μάθησης σε επιστημονικές δημοσιεύσεις (Papers With Code 2024). | 60 |
| 3.20 | Τύποι προβλημάτων αλγορίθμου PPO [Papers With Code (2024)]. | 60 |
| figno:1 | | 64 |

1 Εισαγωγή

1.1 Στόχος

Η παρούσα διπλωματική εργασία πραγματεύεται την εκπαίδευση πρακτόρων τεχνητής νοημοσύνης, προκειμένου να μάθουν να επιτελούν μία εργασία (*task*), σε περιβάλλον παιχνιδιού. Η εκπαίδευση γίνεται χρησιμοποιώντας διαφορετικούς αλγορίθμους Μηχανικής Μάθησης (*Machine Learning*). Τελικός στόχος αποτελεί η σύγκριση των αλγορίθμων αυτών, η οποία πραγματοποείται σε 2 άξονες:

- όσον αφορά τον **χρόνο εκπαίδευσης**, δηλ. τον χρόνο που απαιτήθηκε για να πετύχει ο πράκτορας την καλύτερη επίδοση του στο παιχνίδι και
- όσον αφορά την **τελική επίδοση**, δηλ. το πόσο καλά επιτελεί ο εκπαιδευμένος πράκτορας την εργασία του στο παιχνίδι.

Αναλυτικότερα, το περιβάλλον παιχνιδιού που αναπτύχθηκε είναι ένας χώρος parking και η εργασία που ο πράκτορας καλείται να μάθει είναι η στάθμευση ενός αυτοκινήτου σε μία συγκεκριμένη, αλλά τυχαία θέση στον χώρο αυτό. Επελέγη η συγκεκριμένη εργασία μετά από αρκετή σκέψη, καθώς εκτός του ενδιαφέροντος που παρουσιάζει, προσφέρει και τη δυνατότητα μελέτης ενός προβλήματος με εφαρμογές στον πραγματικό κόσμο. Πράγματι, η στάθμευση αποτελεί ένα πρόβλημα που καλούνται να αντιμετωπίσουν καθημερινά τα αυτόνομα οχήματα, τα οποία πρέπει να είναι σε θέση να επιτελούν αυτή την εργασία με ασφάλεια, ταχύτητα και αποτελεσματικότητα. Επομένως, η επιτυχής επίλυση του προβλήματος της στάθμευσης μέσω της Μηχανικής Μάθησης, ακόμα και σε περιβάλλον προσομοίωσης, αποτελεί ένα πρώτο βήμα προς την κατεύθυνση της ανάπτυξης αυτόνομων οχημάτων.

Ακόμα, σε προσωπικό επίπεδο και ως αρχάριος σε θέματα Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης, στόχος μου μέσα από την εκπόνηση της παρούσας εργασίας, αποτέλεσε η πρώτη εξοικείωση με το αντικείμενο και η κατανόηση των βασικών αρχών και τρόπων λειτουργίας συστημάτων Τεχνητής Νοημοσύνης. Έχοντας πλέον ολοκληρώσει τη διπλωματική μου εργασία, θεωρώ πως ο στόχος αυτός επετεύχθη και μάλιστα ευχάριστα, με τον ψυχαγωγικό τρόπο που προσφέρει η ενασχόληση με ένα παιχνίδι.

Ωστόσο, κατά την εκπόνηση της παρούσας εργασίας, αντιμετώπισα αρκετά προβλήματα και υπέπεισα σε λάθη, τα οποία μου κόστισαν σε χρόνο και ενέργεια. Η αναζήτηση πληροφοριών στο

1.2 Διάρθρωση Διπλωματικής Εργασίας

διαδίκτυο είναι μία χρονοβόρα διαδικασία, εξαιτίας του τεράστιου όγκου διαθέσιμης πληροφορίας. Κάποιος άπειρος μπορεί εύκολα να χαθεί στην πληθώρα διαφορετικών πηγών και επιστημονικών δημοσιεύσεων και να σπαταλήσει χρόνο σε πληροφορίες μη χρήσιμες για αυτόν ή υπερβολικά λεπτομερείς.

Για αυτό τον λόγο, μέσα από το κείμενο αυτής της διπλωματικής εργασίας, επιθυμώ να βοηθήσω νέους ερευνητές του πεδίου να εισαχθούν πιο ομαλά σε αυτό. Στο πνεύμα αυτό, θα προσπαθήσω να εξηγώ τα θέματα της παρούσας εργασίας με τρόπο απλό και σαφή, ώστε να μην χρειάζεται κάποιος να έχει προηγούμενη εμπειρία για να τα κατανοήσει. Έτσι, ελπίζω αυτή η διατριβή να χρησιμοποιηθεί ως αφετηρία από μελλοντικούς φοιτητές με παρόμοιο αντικείμενο μελέτης και η αναγνώση της δικής μου εμπειρίας να τους αποτρέψει από την επανάληψη των ίδιων λαθών.

1.2 Διάρθρωση Διπλωματικής Εργασίας

2 Επισκόπηση του χώρου

Σε αυτό το Κεφάλαιο, θα πραγματοποιήσουμε μία επισκόπηση του χώρου που επιλέχθηκε ως αντικείμενο μελέτης της παρούσας διπλωματικής εργασίας, δηλ. του χώρου της αυτόματης στάθμευσης και γενικότερα, της αυτόνομης οδήγησης. Αρχικά, στην Ενότητα 2.1 θα παρουσιάσουμε τον τρόπο λειτουργίας και την τρέχουσα κατάσταση της τεχνολογίας των αυτόνομων οχημάτων στον πραγματικό κόσμο. Στη συνέχεια, στην Ενότητα 2.2 θα αναφερθούμε σε προηγούμενες εργασίες και επιστημονικές δημοσιεύσεις που ασχολήθηκαν με την αυτόματη στάθμευση σε περιβάλλοντα προσομοιώσεων και παιχνιδιών.

2.1 Αυτόνομα οχήματα

2.1.1 Ορισμός

Ως «αυτόνομο όχημα», ορίζεται η τεχνολογία για μερική ή πλήρη αντικατάσταση του ανθρώπινου οδηγού στην πλοϊγηση ενός οχήματος από την αφετηρία στον προορισμό, αποφεύγοντας κινδύνους του δρόμου και ανταποκρινόμενη στις κυκλοφοριακές συνθήκες (University of Michigan Center for Sustainable Systems 2023). Τα αυτόνομα οχήματα χωρίζονται σε 6 επίπεδα αυτοματισμού, με βάση το βαθμό στον οποίο ένα όχημα μπορεί να λειτουργήσει χωρίς ανθρώπινη παρέμβαση, σύμφωνα με τον οργανισμό SAE International (Society of Automobile Engineers 2021). Τα 6 αυτά επίπεδα φαίνονται στην Εικόνα 2.1.

Τα πρώτα 3 επίπεδα (0-2) αναφέρονται σε οχήματα που απαιτούν την παρουσία ενός ανθρώπου οδηγού, αλλά χρησιμοποιούν αυτοματισμούς για την ασφάλεια, όπως προειδοποιήσεις για τυφλά σημεία και αυτόματο φρενάρισμα. Τα επίπεδα 3 και 4 αντιπροσωπεύουν τεχνολογία στην οποία το όχημα είναι αυτόνομο κάτω από συγκεκριμένες συνθήκες, ενώ κάτω από άλλες συνθήκες χρειάζεται η παρέμβαση ενός ανθρώπου. Τέλος, το επίπεδο 5 είναι το μόνο επίπεδο στο οποίο ένα όχημα θεωρείται πλήρως αυτόνομο και δεν απαιτείται ποτέ ανθρώπινη παρέμβαση.

2.1.2 Αρχές λειτουργίας

Τα αυτόνομα οχήματα στηρίζονται κυρίως στην τεχνητή νοημοσύνη και απαιτούν μεγάλο πλήθος δεδομένων για την εκπαίδευση τους. Τα δεδομένα αυτά συλλέγονται από αισθητήρες, όπως κάμερες,

2.1 Αυτόνομα οχήματα

SAE J3016™ LEVELS OF DRIVING AUTOMATION™
Learn more here: sae.org/standards/content/j3016_202104

Copyright © 2021 SAE International. The summary table may be freely copied and distributed AS-IS provided that SAE International is acknowledged as the source of the content.

| SAE LEVEL 0™ | SAE LEVEL 1™ | SAE LEVEL 2™ | SAE LEVEL 3™ | SAE LEVEL 4™ | SAE LEVEL 5™ |
|--|---|--|---|---|--|
| What does the human in the driver's seat have to do? | You are driving whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering | You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety | You are not driving when these automated driving features are engaged – even if you are seated in “the driver’s seat” | When the feature requests, you must drive | These automated driving features will not require you to take over driving |
| What do these features do? | These features are limited to providing warnings and momentary assistance | These features provide steering OR brake/acceleration support to the driver | These features provide steering AND brake/acceleration support to the driver | These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met | This feature can drive the vehicle under all conditions |
| Example Features | <ul style="list-style-type: none"> • automatic emergency braking • blind spot warning • lane departure warning | <ul style="list-style-type: none"> • lane centering OR • adaptive cruise control | <ul style="list-style-type: none"> • lane centering AND • adaptive cruise control at the same time | <ul style="list-style-type: none"> • traffic jam chauffeur | <ul style="list-style-type: none"> • local driverless taxi • pedals/steering wheel may or may not be installed |
| | | | | | <ul style="list-style-type: none"> • same as level 4, but feature can drive everywhere in all conditions |

Copyright © 2021 SAE International.

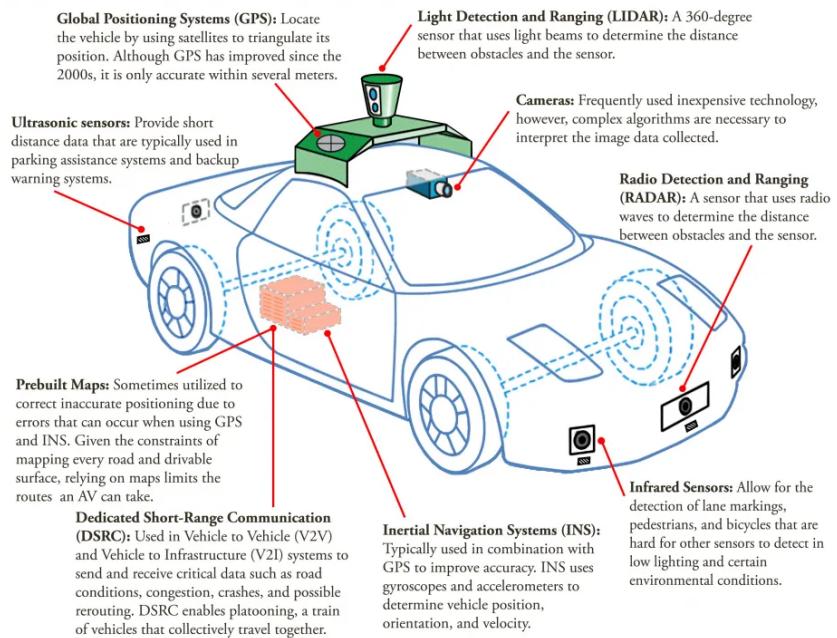
Εικόνα 2.1. Επίπεδα αυτοματισμού αυτόνομων οχημάτων (Society of Automobile Engineers 2021).

ραντάρ και λέιζερ (*LIDAR - Light Detection and Ranging*). Οι κάμερες χρησιμοποιούν Μηχανική Όραση (*Computer Vision*) για να αναγνωρίσουν τα αντικείμενα, ενώ τα ραντάρ και τα LIDAR χρησιμεύουν για την ανίχνευση της απόστασης και της ταχύτητας των αντικειμένων. Πιο συγκεκριμένα, η τεχνολογία LIDAR εκπέμπει μία ταχεία αλληλουχία από πολύ μικρούς παλμούς λέιζερ και μετρά το χρόνο που χρειάζεται για να επιστρέψουν από τα αντικείμενα που βρίσκονται στο δρόμο. Από το χρόνο αυτό, μπορεί να υπολογιστεί η απόσταση των αντικειμένων από το όχημα, ενώ από τη διαφορά των χρόνων μεταξύ διαδοχικών παλμών αντλούνται πληροφορίες για το σχήμα του κάθε αντικειμένου. Μάλιστα, χρησιμοποιώντας τεχνολογίες Ενσωματωμένης Φωτονικής (*Integrated Photonics*), όπως διαμορφωτές Mach-Zender και ανιχνευτές φωτός, η ακρίβεια στην ανάλυση των σχημάτων μπορεί να φτάσει το 1mm, κάτι που ξεπερνάει την ανθρώπινη όραση και αντίληψη (Saini 2019). Επομένως, με αυτούς τους τρόπους τα αυτόνομα οχήματα είναι σε θέση να εντοπίζουν και να αναγνωρίζουν τα στοιχεία του περιβάλλοντος οδήγησης, όπως φανάρια, δέντρα, πεζούς, πινακίδες κυκλοφορίας κλπ. Στη συνέχεια, τα δεδομένα που συλλέγονται από τους αισθητήρες στέλνονται στο λογισμικό του οχήματος, όπου μέσω νευρωνικών δικτύων και αλγορίθμων μηχανικής μάθησης λαμβάνονται οι αποφάσεις για την κίνηση του.

Ακόμα, πολλές φορές χρησιμοποιείται η τεχνολογία *Geofencing* για να βοηθήσει στην πλοήγηση των αυτόνομων οχημάτων. Η τεχνολογία αυτή βασίζεται στο σύστημα Global Positioning System (*GPS*) για να δημιουργήσει εικονικά όρια (*geofences*) σε μία συγκεκριμένη γεωγραφική περιοχή. Τα εικονικά όρια χρησιμοποιούνται για να ενεργοποιήσουν αυτόματες ενέργειες ή ειδοποιήσεις όταν

ένα όχημα εισέρχεται ή εξέρχεται από αυτά. Με αυτόν τον τρόπο, τα αυτόνομα οχήματα μπορούν να αναγνωρίσουν την περιοχή στην οποία κινούνται και να προσαρμόσουν την κίνησή τους ανάλογα (Gillis 2024).

Οι παραπάνω τεχνολογίες των αυτόνομων οχημάτων παρουσιάζονται στην Εικόνα 2.2.



Εικόνα 2.2. Τεχνολογίες αυτόνομων οχημάτων (University of Michigan Center for Sustainable Systems 2023)

2.1.3 Πλεονεκτήματα και Μειονεκτήματα

Η ευρεία υιοθέτηση των αυτόνομων οχημάτων θα έχει θετικές και αρνητικές επιπτώσεις στην κοινωνία, οι οποίες περιγράφονται αναλυτικά και με παραστατικό τρόπο στο (Bright Side 2020). Τα σημαντικότερα πλεονεκτήματα και μειονεκτήματα της χρήσης των αυτόνομων οχημάτων παρατίθενται παρακάτω.

Στα πλεονεκτήματα περιλαμβάνονται:

- η μείωση των τροχαίων ατυχημάτων. Σύμφωνα με το (Singh 2015), περισσότερο από 90% των τροχαίων ατυχημάτων προκαλούνται από τον ανθρώπινο παράγοντα όπως απόσπαση προσοχής, κακή λήψη αποφάσεων ή κατανάλωση αλκοόλ. Επομένως, η αυτοματοποίηση της οδήγησης μπορεί να μειώσει σημαντικά τον αριθμό των ατυχημάτων.
- η δυνατότητα μετακίνησης για άτομα που δεν είναι ικάνα να οδηγήσουν, όπως ηλικιωμένοι ή άτομα με αναπηρία.

2.1 Αυτόνομα οχήματα

- η μείωση της κυκλοφοριακής συμφόρησης. Αυτό είναι εφικτό χάρη στην επικοινωνία μεταξύ των αυτόνομων οχημάτων και το συντονισμό τους.
- η μείωση της περιβαλλοντικής ρύπανσης. Τα αυτόνομα οχήματα πετυχαίνουν πιο αποδοτική οδήγηση, η οποία, σύμφωνα με μελέτη του πανεπιστημίου του Michigan (Erickson 2018), μπορεί να οδηγήσει σε μείωση των εκπομπών ρύπων κατά 9% σε σύγκριση με τα συμβατικά οχήματα.
- η αύξηση του ελεύθερου χρόνου των ανθρώπων, καθώς δεν θα χρειάζεται να έχουν την προσοχή τους στην οδήγηση.
- η διευκόλυνση της στάθμευσης. Τα αυτόνομα οχήματα μπορούν να αφήνουν τους επιβάτες τους απευθείας στον προορισμό τους κι έπειτα να αναζητούν χώρο στάθμευσης και να παρκάρουν αυτόματα.

Στα μειονεκτήματα περιλαμβάνονται:

- η μείωση θέσεων εργασίας. Η ευρεία χρήση των αυτόνομων οχημάτων ενδέχεται να επηρεάσει εργαζόμενους στον τομέα της οδήγησης όπως οδηγούς ταξί και φορτηγών.
- η αύξηση του κόστους. Η απαραίτητη τεχνολογία σε υλικό και σε λογισμικό για τη λειτουργία των αυτόνομων οχημάτων μπορεί να αυξήσει σημαντικά το κόστος τους σε σχέση με τα συμβατικά οχήματα.
- η έλλειψη νομοθεσίας. Η ανάπτυξη των αυτόνομων οχημάτων αντιμετωπίζει πληθώρα νομικών προβλημάτων που προκύπτουν από τη χρήση τους. Για παράδειγμα, ποιος είναι υπεύθυνος σε περίπτωση ατυχήματος; Ο ιδιοκτήτης του οχήματος, ο κατασκευαστής του ή ο προγραμματιστής του λογισμικού;
- το ηλεκτρονικό έγκλημα. Το λογισμικό των αυτόνομων οχημάτων θα αποτελέσει στόχο κακόβουλων προγραμματιστών (hackers) που θα επιδιώκουν να κλέψουν προσωπικά δεδομένα των χρηστών ή ακόμα και να ανακατευθύνουν τα οχήματα ή να προκαλέσουν ατυχήματα .

2.1.4 Η κατάσταση σήμερα

Παρόλο που πολλοί ειδικοί του χώρου προέβλεπαν ότι τα αυτόνομα οχήματα θα κυκλοφορούσαν ήδη στους δρόμους, όπως ο ιδρυτής της Tesla Elon Musk, ο οποίος υποστήριζε πως από το 2020 τα αυτοκίνητα της εταιρίας του θα ήταν πλήρως αυτόνομα και δεν θα απαιτούσαν την προσοχή των επιβατών τους (Trudell 2024), η πραγματικότητα είναι διαφορετική.

Σήμερα, τα συστήματα αυτόματης οδήγησης προς πώληση ανήκουν στο επίπεδο 2, όπως το Super Cruise της εταιρίας Grand Motors και το Full Self Driving της εταιρίας Tesla. Τα συστήματα αυτά έχουν χαρακτηριστικά όπως το σύστημα προσαρμοστικού ρυθμού (*Adaptive Cruise Control*) για την αυτόματη ρύθμιση της ταχύτητας, το σύστημα παρακολούθησης λωρίδας (*Lane Keeping*) για τη διατήρηση του οχήματος στη λωρίδα του και το σύστημα αυτόματης στάθμευσης (*Autopark*) για τη μετακίνηση σε παράλληλες ή κάθετες θέσεις στάθμευσης (Tucker 2024). Ωστόσο, εξακολουθεί να είναι απαραίτητη η συνεχής προσοχή και ετοιμότητα του οδηγού.

Υπάρχουν ακόμα πολλά εμπόδια που καθυστερούν την ανάπτυξη και εμπορική χρήση πλήρως αυτόνομων οχημάτων. Κάποια από αυτά αναφέρθηκαν νωρίτερα στα μειονεκτήματα τους, όπως η ανάγκη θέσπισης σχετικής νομοθεσίας, η διασφάλιση της ασφάλειας των δεδομένων και η ζητούμενη μείωση του κόστους παραγωγής των τεχνολογιών αυτόνομης οδήγησης. Επίσης, πρόκληση αποτελεί το γεγονός ότι τα αυτόνομα οχήματα δεν μπορούν να λειτουργήσουν κάτω από κακές καιρικές συνθήκες, καθώς οι αισθητήρες τους εμποδίζονται από τη βροχή, το χιόνι ή την ομίχλη. Επιπλέον, η κακή κατάσταση των δρόμων, με παραδείγματα όπως λακούβες και ελλιπή σήμανση, δυσκολεύει την κατανόηση του περιβάλλοντος οδήγησης από τα αυτόνομα οχήματα. Όλοι οι παραπάνω περιορισμοί πρέπει να αρθούν ώστε τα αυτόνομα οχήματα να γίνουν πραγματικά ασφαλή, να κερδίσουν την εμπιστοσύνη των καταναλωτών και να επιτύχουν ευρεία υιοθέτηση.

Δυστυχώς, μεγάλη μερίδα του κόσμου δεν καταλαβαίνει ότι τα σημερινά συστήματα παρέχουν μόνο μερικώς αυτόματη οδήγηση. Αυτό οφείλεται κυρίως στην παραπλανητική προώθηση (*Marketing*) από εταιρίες όπως η Tesla, η οποία με τον όρο «Full Self Driving» δημιουργεί την εντύπωση ότι τα αυτοκίνητά της είναι πλήρως αυτόνομα. Αυτό έχει ως αποτέλεσμα την υπερβολική χαλάρωση των οδηγών και την αύξηση των ατυχημάτων. Σύμφωνα με δεδομένα της Εθνική Υπηρεσία Οδικής Ασφάλειας των ΗΠΑ (*NHTSA*), από το 2019 έχουν διερευνηθεί 736 ατυχήματα που συνέβησαν σε αυτοκίνητα της Tesla που ήταν σε λειτουργία αυτόματης οδήγησης, ενώ από αυτά προέκυψαν 17 θάνατοι (*Blanco 2023*).

Πλέον, οι ειδικοί είναι πιο μετριοπαθείς στις προβλέψεις τους, εκτιμώντας πως θα χρειαστούν ακόμα αρκετές δεκαετίες για να επιτευχθεί το επίπεδο 5, δηλαδή το πλήρως αυτόνομο όχημα. Σε μία έρευνα του 2023 από τη διεθνή εταιρεία παροχής συμβουλευτικών υπηρεσιών *McKinsey and Company* (*Deichmann κ.ά. 2023*), προβλέπεται ότι το 2030, το 12% των νέων επιβατικών αυτοκινήτων θα πωλούνται με τεχνολογίες αυτόνομης οδήγησης επιπέδου 3 ή μεγαλύτερου, ενώ το 2035 το 37% αυτών θα έχουν προηγμένες τεχνολογίες αυτόνομης οδήγησης.

2.2 Προηγούμενη έρευνα

Στον τομέα της τεχνητής νοημοσύνης των αυτόνομων οχημάτων, έχουν πραγματοποιηθεί πολλές εργασίες και έρευνες τα τελευταία χρόνια, σε περιβάλλοντα προσομοιώσεων. Οι περισσότερες από αυτές, εστιάζουν στο κομμάτι της οδήγησης, όπως η εργασία των (*Anzalone κ.ά. 2022*), στην οποία χρησιμοποιείται ο αλγόριθμος PPO και η τεχνική της Κλιμακωτής Μάθησης (*Curriculum Learning*), για την εκπαίδευση ενός πράκτορας τεχνητής νοημοσύνης σε σενάρια αυτόματης οδήγησης. Άλλο παράδειγμα αποτελεί το άρθρο των (*Loiacono κ.ά. 2010*), στο οποίο χρησιμοποιείται ο αλγόριθμος Q-Learning για την εκπαίδευση ενός πράκτορα στην προσπέραση άλλων οχημάτων.

Συχνά, τα τελικά αποτελέσματα τέτοιου είδους εργασίων καταγράφονται σε video και αναρτώνται σε πλατφόρμες όπως το YouTube. Αυτό συνηθίζεται, καθώς η κίνηση του πράκτορα δεν μπορεί να

2.2 Προηγούμενη έρευνα

αναπαρασταθεί σε ένα άρθρο, ενώ το video παρέχει μια πιο παραστατική εικόνα της λειτουργίας του αλγορίθμου στην πράξη. Έτσι, γίνεται ευκολότερα κατανοητό στον άνθρωπο το επίπεδο οδήγησης που έχει επιτύχει ο πράκτορας. Επίσης, τα video αυτού του είδους είναι μία καλή εισαγωγή στον χώρο της Ενισχυτικής Μάθησης, επειδή εξηγούν με σύντομο και απλό τρόπο τις βασικές αρχές του πεδίου και τη διαδικασία εκπαίδευσης του πράκτορα. Για αυτό, προτείνω στους αρχάριους του αντικειμένου να παρακολουθήσουν κάποια από αυτά τα video, προτού ασχοληθούν με την ανάγνωση τεχνικών άρθρων. Κάποια αξιοσημείωτα video είναι τα (Porro 2020), (NeuralNine 2021), (Saravia 2020), στα οποία χρησιμοποιείται ο γενετικός αλγόριθμος NEAT και τα (Code-Bullet 2019), (AI-Forge 2022), στα οποία χρησιμοποιούνται οι αλγόριθμοι Q-Learning και PPO αντίστοιχα. Στα παραπάνω video, πρώτα δημιουργείται ένα παιχνίδι αγώνων αυτοκινήτων (*Car Racing*), στο οποίο ο πράκτορας εκπαιδεύεται να ολοκληρώνει τον γύρο της πίστας, όσο το δυνατόν πιο γρήγορα και χωρίς να παρεκτρέπεται εκτός δρόμου. Μετά από αρκετά βήματα εκπαίδευσης, οι πράκτορες φτάνουν σε ικανοποιητικό επίπεδο οδήγησης.

Παράλληλα, υπάρχουν και εργασίες που επικεντρώνονται στο κομμάτι της αυτόματης στάθμευσης (*Car Parking*), αν και είναι πολύ λιγότερες σε σχέση με τις εργασίες οδήγησης. Για παράδειγμα, η εργασία (Lai 2018) χρησιμοποιεί τον αλγόριθμο Q-Learning για την εκπαίδευση ενός πράκτορα στη στάθμευση σε παράλληλη θέση. Ενδιαφέρουσα αποτελεί επίσης η εργασία των (Mujumdar, Shah, και Cui 2020), στην οποία ο πράκτορας καλείται να φτάσει σε μία από τις διαθέσιμες θέσεις στάθμευσης, χωρίς να συγκρουστεί με άλλα οχήματα ή πεζούς και χρησιμοποιείται ο αλγόριθμος PPO καθώς και ο αλγόριθμος μίμησης GAIL. Ακόμα, η εργασία των (Moreira 2021) διεξάγει σύγκριση των αλγορίθμων DDPG, SAC και TD3 για την εργασία της αυτόματης στάθμευσης. Τέλος, ξεχωριστή αναφορά αξίζει το video του (Arzt 2019), καθώς ήταν αυτό που μου κίνησε πρώτο το ενδιαφέρον και με ενέπνευσε να ασχοληθώ με το θέμα της αυτόματης στάθμευσης.

3 Βασική Θεωρία

3.1 Τεχνητή Νοημοσύνη

3.1.1 Ορισμός

Οι Stuart Russell και Peter Norvig, στο βιβλίο τους “Artificial Intelligence: A Modern Approach” ορίζουν την τεχνητή νοημοσύνη ως «το επιστημονικό πεδίο το οποίο αναπτύσσει και μελετά λογισμικό και μεθόδους που επιτρέπουν στις μηχανές να αντιλαμβάνονται το περιβάλλον τους και να χρησιμοποιούν τη μάθηση και τη νοημοσύνη για να επιλέξουν δράσεις που μεγιστοποιούν τις πιθανότητες επίτευξης καθορισμένων στόχων» (Russell και Norvig 2021). Με άλλα λόγια, η τεχνητή νοημοσύνη είναι η τεχνολογία που δίνει τη δυνατότητα στους υπολογιστές και τις μηχανές να προσομοιώνουν την ανθρώπινη νοημοσύνη και να αποκτούν δυνατότητες επίλυσης προβλημάτων. Το ευρύ πεδίο της τεχνητής νοημοσύνης περιλαμβάνει πολλές διαφορετικές επιστήμες, από την επιστήμη των υπολογιστών (*Computer Science*), την ανάλυση δεδομένων (*Data Analytics*), το υλικό (*Hardware*), τη νευροεπιστήμη (*Neuroscience*), έως τη γλωσσολογία (*Linguistics*) και τη φιλοσοφία (*Philosophy*).

3.1.2 Μοντέλα και Πράκτορες τεχνητής νοημοσύνης

Το τελικό αποτέλεσμα μίας εφαρμογής τεχνητής νοημοσύνης συχνά καλείται μοντέλο (*Model*) ή πράκτορας (*Agent*), με τους όρους αυτούς να χρησιμοποιούνται εναλλάξιμα. Ωστόσο, υπάρχει λεπτή διάκριση μεταξύ τους, καθώς έχουν μοναδικές λειτουργίες κι εξυπηρετούν διαφορετικούς σκοπούς (Ezeokeke 2024).

Συγκεκριμένα, τα μοντέλα τεχνητής αναφέρονται σε συστήματα που εκπαιδεύονται σε συγκεκριμένα σύνολα δεδομένων και ως εκ τούτου, είναι συνήθως περιορισμένα στην εκτέλεση συγκεκριμένων εργασιών. Αποτελούν τον “εγκέφαλο” πολλών εφαρμογών τεχνητής νοημοσύνης, όπως τα μοντέλα αναγνώρισης εικόνων για την ανίχνευση προσώπων ή τα γλωσσικά μοντέλα που χρησιμοποιούνται στη μετάφραση κειμένου.

Αντίθετα, οι πράκτορες τεχνητής νοημοσύνης είναι πιο πολύπλοκες οντότητες που όχι μόνο χρησιμοποιούν μοντέλα τεχνητής νοημοσύνης, αλλά αλληλεπιδρούν με το περιβάλλον τους για

3.1 Τεχνητή Νοημοσύνη

να επιτύχουν τους στόχους τους. Συνεπώς, πρόκειται για συστήματα που χαρακτηρίζονται από αυτονομία και είναι σε θέση να ενεργούν ανεξάρτητα. Χρησιμοποιούνται σε εφαρμογές που απαιτείται η λήψη αποφάσεων σε πραγματικό χρόνο, όπως οι εικονικοί βοηθοί, τα αυτόνομα οχήματα και οι ρομποτικές εφαρμογές.

Επομένως, τα μοντέλα τεχνητής νοημοσύνης μπορούν να παρομοιαστούν με μία ισχυρή μηχανή, η οποία όμως χρειάζεται έναν ικανό χειριστή, τον πράκτορα, για να μετατρέψει την πληροφορία της σε δράση.

3.1.3 Ιστορική Εξέλιξη

Ο Άλαν Τούρινγκ ήταν ο πρώτος που διεξήγαγε σημαντικές έρευνες στον τομέα που αυτός ονόμασε «Μηχανική νοημοσύνη» (*Machine Intelligence*). Το 1950, ο Τούρινγκ δημοσιεύει το άρθρο “Υπολογιστικά Μηχανήματα και νοημοσύνη”, στο οποίο θέτει το ερώτημα: “Μπορούν οι μηχανές να σκέφτονται; (*Turing 1950*)” Για να απαντήσει αυτό το ερώτημα,, προτείνει ένα τεστ, γνωστό ως «*Turing Test*». Το τεστ αυτό περιλαμβάνει έναν ανθρώπινο ανακριτή που κάνει ερωτήσεις σε δύο συμμετέχοντες, έναν άνθρωπο και έναν υπολογιστή, χωρίς να γνωρίζει κάθε φορά ποιός απαντάει. Ο στόχος είναι μέσα από τις απαντήσεις τους και εντός ενός χρονικού πλαισίου, ο ανακριτής να μπορέσει να διακρίνει ποιος είναι ο άνθρωπος και ποιος ο υπολογιστής. Έτσι, η ικανότητα του υπολογιστή να σκέπτεται προκύπτει από την πιθανότητα να τον ανοιγνωρίσει εσφαλμένα ο ανακριτής ως άνθρωπο. Μάλιστα, ο Τούρινγκ υποστηρίζει πως εάν ο υπολογιστής αντιδρά και συμπεριφερέται σαν νοήμων ον, τότε έχει συνείδηση. Αυτό το τεστ αποτελεί σημαντικό μέρος της ιστορίας της τεχνητής νοημοσύνης, καθώς και φιλοσοφική συζήτηση, η οποία επανήλθε στο προσκήνιο το 2022, με την κυκλοφορία του ChatGPT, με μερικούς να υποστηρίζουν πως πλέον έχουν επιτευχθεί τα κριτήρια του *Turing Test*.

Ο όρος τεχνητή νοημοσύνη» χρησιμοποιήθηκε για πρώτη φορά το 1956 από τον Τζον Μακάρθι, ο οποίος αναγνωρίζεται ευρέως ως ο πατέρας της τεχνητής νοημοσύνης, λόγω της προσφοράς του στο αντικείμενο. Τότε, στο πρώτο συνέδριο τεχνητής νοημοσύνης στο Πανεπιστήμιο Dartmouth, ο Τζον Μακάρθι όρισε την τεχνητή νοημοσύνη ως την «επιστήμη και μεθοδολογία της δημιουργίας νοημόνων μηχανών» (*McCarthy 2004*).

Την δεκαετία του 1980, τα νευρωνικά δίκτυα που χρησιμοποιούν τον αλγόριθμο της οπισθοδιάδοσης (*Backpropagation*) για την εκπαίδευσή τους, ξεκινάνε να χρησιμοποιούνται ευρέως σε εφαρμογές τεχνητής νοημοσύνης. Η τεχνολογία αυτή αποτελεί τη βάση πολλών συστημάτων τεχνητής νοημοσύνης που χρησιμοποιούνται σήμερα.

Η δεκαετία του 1990 σηματοδοτεί την έναρξη της εποχής της επιτυχίας των συστημάτων τεχνητής νοημοσύνης στην επίλυση προβλημάτων που απαιτούν ανθρώπινη νοημοσύνη. Αρχικά, το 1995, οι Stuart Russell και Peter Norvig δημοσιεύουν το βιβλίο “Artificial Intelligence: A Modern Approach”

, το οποίο αποτελεί ακόμα και σήμερα, το δημοφιλέστερο εγχειρίδιο της τεχνητής νοημοσύνης παγκοσμίως. Το 1997, ο υπολογιστής Deep Blue της IBM κερδίζει τον τότε παγκόσμιο πρωταθλητή στο σκάκι, Garry Kasparov. Αυτή ήταν η πρώτη φορά που ο τρέχων παγκόσμιος πρωταθλητής ηττήθηκε από έναν υπολογιστή.

Στα πιο σύγχρονα χρόνια, το ενδιαφέρον και η χρηματοδότηση στο πεδίο της τεχνητής νοημοσύνης αυξήθηκαν σημαντικά μετά το 2012, όταν η αύξηση της υπολογιστικής ισχύος οδήγησε σε σημαντικές επιτυχίες στον τομέα της Βαθιάς Μάθησης (*Deep Learning*). Έτσι, το 2015, ο υπερυπολογιστής Minwa της Baidu χρησιμοποιεί ένα ειδικό βαθύ νευρωνικό δίκτυο (*Deep Neural Network*), γνωστό ως συνελικτικό νευρωνικό δίκτυο (*Convolutional Neural Network*), για την αναγνώριση και κατηγοριοποίηση εικόνων και πετυχαίνει υψηλότερο βαθμό ακρίβειας από τον μέσο ανθρώπο. Το 2016, το πρόγραμμα AlphaGo της DeepMind -που επίσης τροφοδοτείται από ένα βαθύ νευρωνικό δίκτυο- κερδίζει τον παγκόσμιο πρωταθλητή στο Go, ένα παιχνίδι στρατηγικής. Η νίκη αυτή είναι εξίσου σημαντική με αυτήν έναντια στον Kasparov, δεδομένου του τεράστιου αριθμού πιθανών κινήσεων που υπάρχουν στο Go (πάνω από 14,5 τρισεκατομμύρια μετά από μόλις τέσσερις κινήσεις!). Το 2023, η εμφάνιση των μεγάλων γλωσσικών μοντέλων (*Large Language Models*), όπως το ChatGPT, προκαλεί μια τεράστια αύξηση στις επιδόσεις της τεχνητής νοημοσύνης και την αξία που έχει για ανθρώπους και επιχειρήσεις.

Στη σημερινή εποχή, η παραγωγική τεχνητή νοημοσύνη (*Generative AI*) μπορεί να μάθει και να συνθέσει όχι μόνο ανθρώπινη γλώσσα, αλλά και άλλους τύπους δεδομένων, συμπεριλαμβανομένων εικόνων, βίντεο, μουσικής, κώδικα λογισμικού και ακόμη και μοριακών δομών. Έτσι, το πεδίο της τεχνητής νοημοσύνης βρίσκεται σήμερα σε άνθιση, με εταιρείες όπως η OpenAI, η Google και η NVIDIA να αναπτύσσουν συνεχώς νέες, πρωτοποριακές εφαρμογές και υπηρεσίες που βασίζονται σε αυτήν.

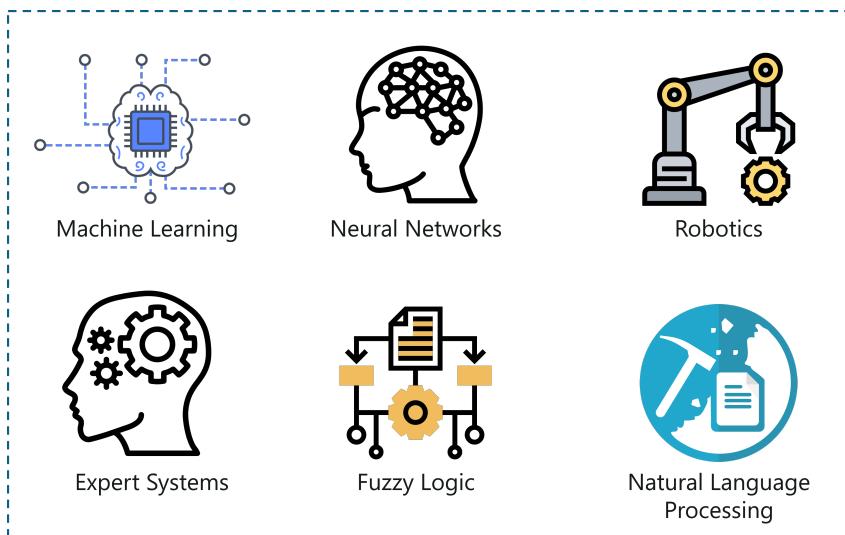
3.1.4 Κατηγορίες

Η τεχνητή νοημοσύνη αποτελεί ένα ευρύ πεδίο, το οποίο περιλαμβάνει πολλές διαφορετικές τεχνολογίες και εφαρμογές. Οι κύριοι κλάδοι της τεχνητής νοημοσύνης παρουσιάζονται στην *Εικόνα 3.1* και περιγράφονται συνοπτικά παρακάτω.

Μηχανική μάθηση

Αποτελεί τον κλάδο της τεχνητής νοημοσύνης που εστιάζει στην ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μαθαίνουν από μεγάλα σύνολα δεδομένων και να κάνουν προβλέψεις ή να παίρνουν αποφάσεις. Τα μοντέλα μηχανικής μάθησης εκπαιδεύονται και βελτιώνουν την απόδοση τους με το χρόνο, χωρίς να χρειάζεται να προγραμματίζονται ρητά για κάθε συγκεκριμένη εργασία. Η μηχανική μάθηση αναλύεται εκτενέστερα στην *Ενότητα 3.2*.

3.1 Τεχνητή Νοημοσύνη



Εικόνα 3.1. Κατηγορίες τεχνητής νοημοσύνης (Lateef 2024).

Περιλαμβάνει τη βαθιά μάθηση, την αυτόματη επεξεργασία γλώσσας, την όραση υπολογιστών, την αναγνώριση φωνής, την αυτόματη προγραμματισμό και την ρομποτική.

Νευρωνικά Δίκτυα (Βαθιά Μάθηση)

Η Βαθιά μάθηση είναι ένας υποκλάδος της μηχανικής μάθησης που χρησιμοποιεί νευρωνικά δίκτυα πολλών επιπέδων -εξού και το βαθιά-, για να προσομοιώσει την πολύπλοκη διαδικασία λήψης αποφάσεων του ανθρώπινου εγκεφάλου. Επομένως, η κύρια διαφορά της βαθιάς μάθησης από την παραδοσιακή μηχανική μάθηση είναι η δομή του νευρωνικού δικτύου. Τα μοντέλα παραδοσιακής μηχανικής μάθησης χρησιμοποιούν απλά νευρωνικά δίκτυα με έως ένα κρυφό επίπεδο. Αντίθετα, τα μοντέλα βαθιάς μάθησης χρησιμοποιούν δύο ή περισσότερα κρυφά επίπεδα (συχνά εκατοντάδες ή χιλιάδες) για να εκπαιδεύσουν τα μοντέλα τους (Johnson 2020). Η εκπαίδευση αυτή απαιτεί μεγάλα σύνολα δεδομένων κι επομένως, σημαντικούς υπολογιστικούς πόρους. Ωστόσο, τα μοντέλα της βαθιάς μάθησης μπορούν να αναγνωρίσουν πολύπλοκα μοτίβα σε δεδομένα όπως είκονες, κείμενα, ήχους, προκειμένου να παράγουν ακριβείς αναλύσεις και προβλέψεις. Έτσι, επιτυγχάνουν να λύσουν πιο πολύπλοκα προβλήματα, όπως η αναγνώριση εικόνων, η μετάφραση γλώσσας και η αυτόνομη οδήγηση. Συνολικά, η βαθιά μάθηση είναι η τεχνολογία που κινητοποιεί τις περισσότερες εφαρμογές της τεχνητής νοημοσύνης στη σημερινή ζωή.

Ρομποτική

Η ρομποτική αποτελεί τον κλάδο που επικεντρώνεται στη δημιουργία ευφυών μηχανών (*Robot*), ικανών να εκτελούν εργασίες στον φυσικό κόσμο. Ενσωματώνει τεχνικές τεχνητής νοημοσύνης για

να επιτρέψει στα ρομπότ να αντιλαμβάνονται το περιβάλλον τους, να λαμβάνουν αποφάσεις και να προβαίνουν σε δράσεις αυτόνομα. Έτσι, τα ρομπότ αποτελούν πράκτορες τεχνητής νοημοσύνης, οι οποίοι αλληλεπιδρούν και προσαρμόζονται στις δυναμικές συνθήκες του πραγματικού κόσμου. Χαρακτηριστικό παράδειγμα τέτοιου ρομπότ αποτελεί το ανθρωποειδές Σοφία, η οποία διακρίνεται για τη ρεαλιστική ανθρώπινη εμφάνισή της, τα εκφραστικά χαρακτηριστικά του προσώπου της και τη δυνατότητα της να αλληλεπιδράσει και να συνομιλήσει με ανθρώπους (Hanson-Robotics 2016).

Ειδικά συστήματα

Το πεδίο των ειδικών συστημάτων της τεχνητής νοημοσύνης περιλαμβάνει την ανάπτυξη υπολογιστικών συστημάτων, που μιμούνται την ικανότητα λήψης αποφάσεων ενός ανθρώπου, ειδικού, σε συγκεκριμένο τομέα. Τέτοια συστήματα χρησιμοποιούν μια βάση γνώσης από γεγονότα και κανόνες (λογική if-then-else), μαζί με ένα μηχανισμό για να εφαρμόσουν τη γνώση αυτή σε νέες καταστάσεις και να επιλύσουν πολύπλοκα προβλήματα. Τα ειδικά συστήματα χρησιμοποιούνται ευρέως σε τομείς όπως η ιατρική διάγνωση, ο χρηματοοικονομικός σχεδιασμός και ανίχνευση κακόβουλου λογισμικού.

Ασαφής λογική

Η ασαφής λογική (*Fuzzy Logic*) είναι το πεδίο της τεχνητής νοημοσύνης που επιτρέπει στα συστήματα να χειρίστούν την αβεβαιότητα και την ασάφεια των δεδομένων. Αυτό επιτυγχάνεται χρησιμοποιώντας βαθμούς αληθείας (*degrees of truth*) έναντι αυστηρών τιμών (αληθής/ψευδής) της δυαδικής λογικής (*binary logic). Έτσι, παρέχει προσαρμοστικότητα που επιτρέπει την μοντελοποίηση περίπλοκων, πραγματικών καταστάσεων. Η ασαφής λογική χρησιμοποιείται σε εφαρμογές που διαχειρίζονται ατελή δεδομένα, όπως τα συστήματα ελέγχου -για παράδειγμα, η ρύθμιση της θερμοκρασίας σε συστήματα κλιματισμού- και η λήψη αποφάσεων -για παράδειγμα, η πρόταση ιατρικών διαγνώσεων.

Επεξεργασία φυσικής γλώσσας

Η επεξεργασία φυσικής γλώσσας (*Natural Language Processing - NLP*) είναι το υποπεδίο της τεχνητής νοημοσύνης που πραγματεύεται την ανάπτυξη τεχνολογιών που επιτρέπουν στους υπολογιστές να αναγνωρίζουν, να κατανοούν και να παράγουν ανθρώπινη γλώσσα. Για να το πετύχει αυτό, συνδυάζει την υπολογιστική γλωσσολογία (*Computational Linguistics*) με τη στατιστική μοντελοποίηση, τη μηχανική μάθηση και τη βαθιά μάθηση. Η έρευνα στην επεξεργασία φυσικής γλώσσας έχει επιφέρει την εποχή της παραγωγικής τεχνητής νοημοσύνης, με την κυκλοφορία των μεγάλων γλωσσικών μοντέλων, όπως το ChatGPT, που μπορούν να μάθουν και να συνθέσουν ανθρώπινη γλώσσα, αλλά και άλλους τύπους δεδομένων όπως εικόνες και βίντεο. Έτσι, η επεξεργασία φυσικής γλώσσας

3.1 Τεχνητή Νοημοσύνη

χρησιμοποιείται σε πολλές εφαρμογές, όπως οι ψηφιακοί βοηθοί, οι μηχανές μετάφρασης, οι μηχανές αναζήτησης, τα συστήματα αναγνώρισης φωνής.

3.1.5 Εφαρμογές

Σήμερα, η τεχνητή νοημοσύνη είναι το επίκεντρο πολλών τεχνολογιών που υποστηρίζουν υπηρεσίες και αγαθά, τα οποία χρησιμοποιούμε καθημερινά. Ορισμένες από τις πιο συνηθισμένες εφαρμογές της τεχνητής νοημοσύνης περιλαμβάνουν:

- **αναγνώριση ομιλίας:** η αυτόματη μετατροπή της προφορικής ομιλίας σε γραπτό κείμενο γίνεται με τεχνικές της επεξεργασίας φυσικής γλώσσας. Χρησιμοποιείται από ψηφιακούς βοηθούς όπως η Siri της Microsoft και η Alexa της Amazon.
- **αναγνώριση εικόνας:** η αναγνώριση και κατηγοριοποίηση εικόνων γίνεται από αλγόριθμους βαθιάς μάθησης, όπως τα συνελικτικά νευρωνικά δίκτυα και έχει εφαρμογές σε τομείς όπως η αυτόνομη οδήγηση και η αναγνώριση προσώπων.
- **αυτόματη μετάφραση:** εφαρμογές όπως το Google Translate χρησιμοποιούν αλγορίθμους βαθιάς μάθησης για να μεταφράσουν κείμενο από μία γλώσσα σε άλλη.
- **συστήματα προτάσεων:** τα συστήματα αυτά χρησιμοποιούν αλγορίθμους μηχανικής μάθησης για να αναλύσουν τα δεδομένα από πλατφόρμες όπως το YouTube, το Spotify και το Netflix και να προτείνουν περιεχόμενο στους χρήστες, με βάση τις προηγούμενες προτιμήσεις τους.
- **εμπορική προώθηση:** η τεχνητή νοημοσύνη χρησιμοποιείται στην ανάλυση δεδομένων από πλατφόρμες κοινωνικών δικτύων και ηλεκτρονικών καταστημάτων, όπως το Facebook και η Amazon, για την πρόβλεψη των προτιμήσεων των καταναλωτών και την προσωποποίηση των διαφημίσεων και των προσφορών.
- **συστήματα αναζήτησης:** οι αλγόριθμοι αναζήτησης όπως το Google Search χρησιμοποιούν τεχνολογίες τεχνητής νοημοσύνης για να παρέχουν ακριβέστερα αποτελέσματα αναζήτησης.
- **αυτόνομη οδήγηση:** τα αυτόνομα οχήματα χρησιμοποιούν μηχανική όραση για την αναγνώριση του περιβάλλοντος και βαθιά νευρωνικά δίκτυα για τη λήψη αποφάσεων κατά την οδήγηση.
- **οικονομική διαχείριση:** αλγόριθμοι τεχνητής νοημοσύνης για την ανάλυση δεδομένων χρησιμοποιούνται στην πρόβλεψη των χρηματιστηριακών αγορών και τη διαχείριση των επενδύσεων. Έτσι, επενδυτικές πλατφόρμες όπως η Betterment προσφέρουν αυτόματες, εξατομικευμένες οικονομικές συμβουλές, βασισμένες σε αλγόριθμους μηχανικής μάθησης.
- **παραγωγή περιεχομένου:** η παραγωγή περιεχομένου από αλγόριθμους βαθιάς μάθησης, όπως το ChatGPT για κείμενο, το DALL-E για εικόνες και το MuseNet για μουσική, επιτρέπει τη δημιουργία πολυμέσων με βάση την υπόδειξη (*prompt*) του χρήστη.

- **εξυπηρέτηση πελατών:** τα ρομπότ συνομιλίας (*chatbots*) χρησιμοποιούν τεχνολογίες επεξεργασίας φυσικής γλώσσας για να παρέχουν πληροφορίες στους χρήστες και να απαντούν στις ερωτήσεις τους. Μάλιστα, χρησιμοποιούν αλγορίθμους μηχανικής μάθησης ώστε να βελτιώνουν την απόδοσή τους μέσω της εμπερίας τους με τους χρήστες. Παραδείγματα ψηφιακών βοηθών αποτελούν το BlueBot της αεροπορικής εταιρείας KLM και το Ask Mercedes της αυτοκινητοβιομηχανίας Mercedes-Benz.
- **ιατρική διάγνωση:** η τεχνητή νοημοσύνη χρησιμοποιείται στην ανάλυση δεδομένων από ασθενείς, όπως εικόνες από ακτινογραφίες και μαγνητικές τομογραφίες, για την αναγνώριση παθολογιών και την πρόβλεψη ασθενειών. Έτσι, η τεχνητή νοημοσύνη συμβάλλει στην πρόληψη, τη διάγνωση και τη θεραπεία ασθενειών, βελτιώνοντας την ποιότητα της υγειονομικής περίθαλψης.
- **σχεδιασμός διαδρομής :** υπηρεσίες όπως το Google Maps χρησιμοποιούν την τεχνητή νοημοσύνη για να αναλύουν τις συνθήκες κυκλοφορίας και να παρέχουν τις γρηγορότερες διαδρομές, βοηθώντας τους οδηγούς να εξικονομήσουν χρόνο και καύσιμα.
- **ανάπτυξη παιχνιδιών:** αλγόριθμοι τεχνητής νοημοσύνης χρησιμοποιούνται για τη δημιουργία χαρακτήρων που δεν ελέγχονται από τον παίκτη (*Non Player Characters*) με ευφυή συμπεριφορά, που περιλαμβάνει την προσαρμογή στις ενέργειες του παίκτη και την αναγνώριση των προτιμήσεών του, ώστε να προσφέρουν μια πιο ρεαλιστική και διασκεδαστική εμπειρία.

3.2 Μηχανική Μάθηση

3.2.1 Ορισμός

Ο όρος «Μηχανική Μάθηση» επινοήθηκε από τον Arthur Samuel το 1959 και περιγράφηκε ως «η ικανότητα ενός υπολογιστή να μαθαίνει χωρίς να προγραμματιστεί ρητά» (Samuel 1959). Ο Tom Mitchell έδωσε το 1997 έναν διάσημο, πιο μαθηματικό ορισμό των αλγορίθμων μηχανικής μάθησης: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία Ε ως προς μια κλάση εργασιών Τ και μέτρο απόδοσης P, αν η απόδοσή του στις εργασίες της κλάσης Τ, όπως μετράται από το μέτρο απόδοσης P, βελτιώνεται με την εμπειρία Ε» (Mitchell 1997).

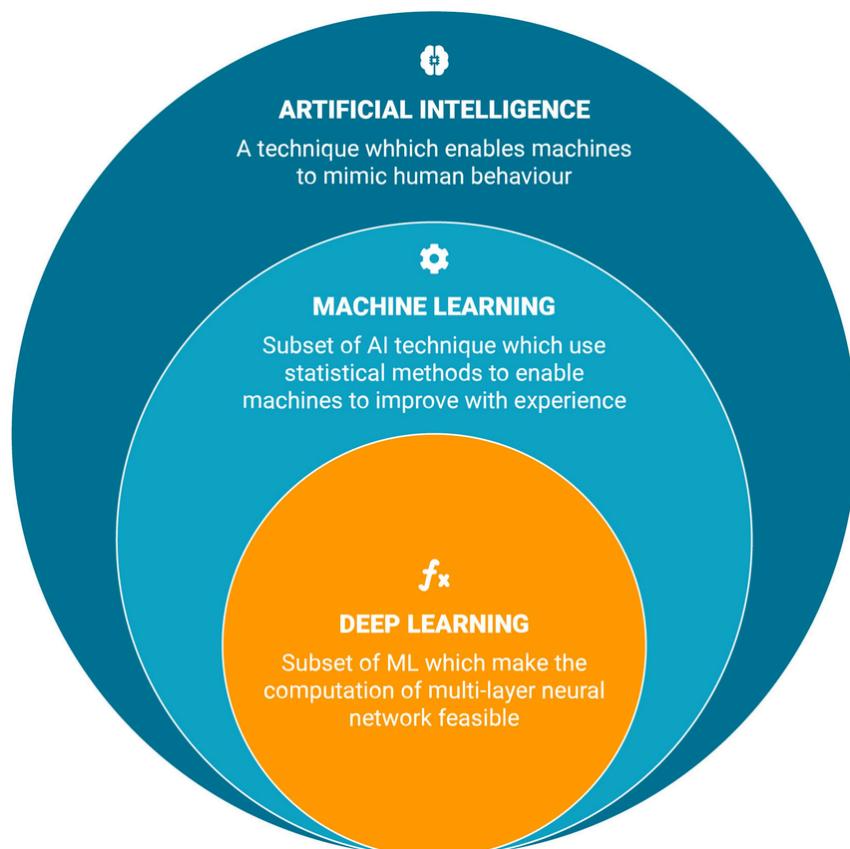
Επομένως, γίνεται σαφές ότι η μηχανική μάθηση ασχολείται με την ανάπτυξη αλγορίθμων, οι οποίοι μαθαίνουν να εκτελούν εργασίες μέσα από δεδομένα ή από προηγούμενη εμπειρία τους και όχι μέσω συγκεκριμένων εντολών. Αυτό είναι πολύ σημαντικό, καθώς τα παραχόμενα μοντέλα μηχανικής μάθησης είναι σε θέση να γενικεύουν τις γνώσεις τους και να τις εφαρμόζουν σε νέα δεδομένα. Αυτό το χαρακτηριστικό τους τα κάνει ιδιαίτερα χρήσιμα για την επίλυση προβλημάτων, τα οποία δεν μπορούν να επιλυθούν με την κλασική προγραμματιστική προσέγγιση. Ωστόσο, η εκπαίδευση των μοντέλων μηχανικής μάθησης απαιτεί μεγάλο όγκο δεδομένων, καθώς και χρόνο και πόρους για την

3.2 Μηχανική Μάθηση

επεξεργασία τους, ενώ το τελικό αποτέλεσμα εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων. Συγκεκριμένα, τα μοντέλα μπορεί να αναπτύξουν προκαταλήψεις (*bias*), εφόσον αυτές υπάρχουν στα δεδομένα εκπαίδευσης τους. Για παράδειγμα, η μελέτη των (Thomson και Thomas 2023), ανακάλυψε κοινωνικές προκαταλήψεις στην εφαρμογή παραγωγής εικόνων Midjourney. Όταν ζητήθηκε η παραγωγή εικόνων ανθρώπων σε εξειδικευμένα επαγγέλματα, τα αποτελέσματα απεικόνιζαν πάντα άνδρες, ενισχύοντας τη φύλετική προκατάληψη του ρόλου των γυναικών στον χώρο εργασίας.

3.2.2 Κατάταξη πεδίου

Σήμερα, πολλοί συγχέουν τους όρους «Τεχνητή Νοημοσύνη» και «Μηχανική Μάθηση». Οφείλει να γίνει κατανοητό, πως η μηχανική μάθηση αποτελεί ένα υποσύνολο της τεχνητής νοημοσύνης, που εστιάζει στη βελτίωση της απόδοσης ενός συστήματος με βάση την εμπειρία. Ορισμένα συστήματα τεχνητής νοημοσύνης χρησιμοποιούν μεθόδους Μηχανικής Μάθησης, ενώ άλλα όχι. Αυτό φαίνεται με παραστατικό τρόπο, στην εικόνα [3.2].

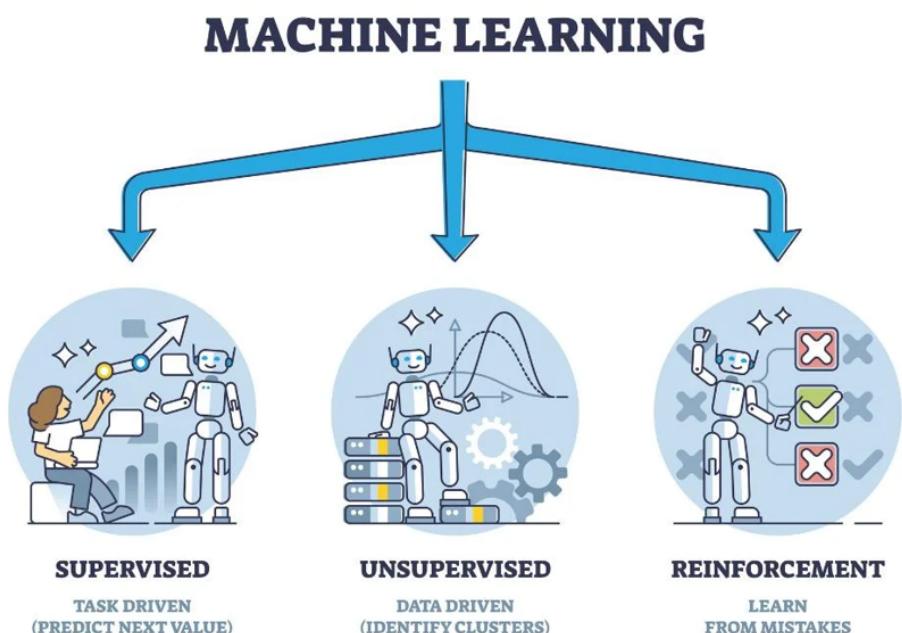


Εικόνα 3.2. Ιεραρχία πεδίων τεχνητής νοημοσύνης (Zand κ.ά. 2022).

Μάλιστα, η εικόνα [3.2] παρουσιάζει και τη σχέση μεταξύ μηχανικής μάθησης και βαθιάς μάθησης, δείχνοντας πως η βαθιά μάθηση αποτελεί μια περαιτέρω εξειδίκευση της μηχανικής μάθησης. Αμφότερες οι τεχνικές αυτές χρησιμοποιούν τεχνητά νευρωνικά δίκτυα για να «μάθουν» από τα δεδομένα. Όμως, η βαθιά μάθηση χρησιμοποιεί πιο πολύπλοκα, πολυεπίπεδα νευρωνικά δίκτυα, τα οποία απαιτούν μεγαλύτερο όγκο δεδομένων και πόρων για την εκπαίδευσή τους.

3.2.3 Κατηγορίες

Η μηχανική μάθηση χωρίζεται σε τρεις κύριες κατηγορίες, ανάλογα με τον τρόπο εκπαίδευσης των μοντέλων: την επιβλεπόμενη μάθηση, την μη επιβλεπόμενη μάθηση και την ενισχυτική μάθηση. Οι τρεις αυτές κατηγορίες παρουσιάζονται στην εικόνα [3.3] και περιγράφονται παρακάτω.



Εικόνα 3.3. Κατηγορίες μηχανικής μάθησης (Stewart 2023).

Επιβλεπόμενη Μάθηση

Η Επιβλεπόμενη Μάθηση (*Supervised Learning*) είναι ο πιο συνηθισμένος τύπος μηχανικής μάθησης. Πήρε το όνομα της καθώς η εκπαίδευση γίνεται υπό επίβλεψη, δηλαδή το μοντέλο μαθαίνει μέσω παραδειγμάτων. Συγκεκριμένα, παρέχεται στο μοντέλο ένα σύνολο δεδομένων εκπαίδευσης με ετικέτες (*labeled data*). Αυτό σημαίνει ότι κάθε δεδομένο αποτελείται από ένα ζεύγος εισόδου-επιθυμητής εξόδου. Το μοντέλο κατά την εκπαίδευση, εντοπίζει μοτίβα στα δεδομένα και προβλέπει

3.2 Μηχανική Μάθηση

για κάθε είσοδο, ποιά είναι η αντίστοιχη έξοδος. Όταν κάνει λάθος, το μοντέλο αναπροσαρμόζεται, μέχρι να μάθει να αντιστοιχίζει σωστά τις εισόδους στις αντίστοιχες εξόδους. Το ζητούμενο είναι, το τελικό μοντέλο να μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις σε νέα δεδομένα -δηλ. δεδομένα που δεν υπήρχαν στο σύνολο εκπαίδευσης- και να πετυχαίνει σε αυτά μεγάλο ποσοστό επιτυχίας. This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples

Κλασικό παράδειγμα επιβλεπόμενης μάθησης αποτελεί η αναγνώριση του φύλου από εικόνες. Στο σύνολο δεδομένων εκπαίδευσης, κάθε εικόνα έχει ετικέτα με το φύλο του ατόμου που απεικονίζεται. Το μοντέλο εκπαιδεύεται να αναγνωρίζει τα χαρακτηριστικά που διαφοροποιούν τα δύο φύλα και να κάνει προβλέψεις για το φύλο του ατόμου. Υπάρχει περίπτωση, το μοντέλο να κάνει σωστές προβλέψεις στις εικόνες που εκπαιδεύτηκε, αλλά όχι σε άγνωστες εικόνες. Τότε, το μοντέλο δεν έχει μάθει να αναγνωρίζει σωστά το φύλο, αλλά απλά έχει μάθει να απαντάει σωστά στα δεδομένα εκπαίδευσης του. Το φαινόμενο αυτό ονομάζεται υπερεκπαίδευση (*overfitting*), αποτελεί ανεπιθύμητη συμπεριφορά και θα αναλυθεί εκτενέστερα σε επόμενη ενότητα.

Επίσης, αξίζει να σημειωθεί πως όταν η έξοδος είναι μία τιμή από ένα πεπερασμένο σύνολο τιμών (όπως πχ πριν, άντρας/γυναίκα), τότε το πρόβλημα μάθησης ονομάζεται Ταξινόμηση (*Classification*). Αντίθετα, όταν η έξοδος είναι μία συνεχής τιμή (πχ η αυριανή θερμοκρασία), τότε το πρόβλημα ονομάζεται Παλινδρόμηση (*Regression*). Συνήθεις αλγόριθμοι που χρησιμοποιούνται για προβλήματα ταξινόμησης είναι τα Δέντρα Απόφασης (*Decision Trees*) και οι Μηχανές Διανυσμάτων Υποστήριξης (*Support Vector Machines*). Αντίθετα, για προβλήματα παλινδρόμησης χρησιμοποιούνται αλγόριθμοι όπως η Γραμμική Παλινδρόμηση (*Linear Regression*) και η Πολυωνυμική Παλινδρόμηση (*Polynomial Regression*).

Μη Επιβλεπόμενη Μάθηση

Η μη επιβλεπόμενη μάθηση (*Unsupervised Learning*) αναφέρεται στις περιπτώσεις όπου τα δεδομένα δεν έχουν ετικέτες, δηλαδή το μοντέλο δεν γνωρίζει την επιθυμητή έξοδο για κάθε δεδομένο. Πλέον, στόχος είναι η αναγνώριση μοτίβο, δομών ή συσχετίσεων στα δεδομένα. Οι πιο συνηθισμένες εργασίες μη επιβλεπόμενης μάθησης είναι η ομαδοποίηση και η μείωση διαστάσεων.

Η ομαδοποίηση (*Clustering*) αφορά την οργάνωση των δεδομένων, χωρίζοντας τα σε ομάδες με παρόμοια χαρακτηριστικά. Ένα παράδειγμα αποτελεί η οργάνωση των πελατών μίας επιχείρησης σε ομάδες με βάση το ιστορικό αγορών τους, έτσι ώστε να εφαρμοστεί διαφορετική στρατηγική προώθησης για κάθε ομάδα. Παραδείγματα αλγορίθμων ομαδοποίησης είναι ο K-Means και η Ιεραρχική Ομαδοποίηση (*Hierarchical Clustering*).

Η μείωση διαστάσεων (*Dimensionality Reduction*) αφορά τη μείωση του αριθμού των χαρακτηριστικών που περιγράφουν τα δεδομένα, διατηρώντας όμως σε μεγάλο βαθμό την πληροφορία που περιέχουν.

Αυτό είναι χρήσιμο για την απλοποίηση μοντέλων, την αφαίρεση θορύβου από τα δεδομένα και την οπτικοποίηση τους. Ένας από τους πιο διάσημους αλγορίθμους μείωσης διαστάσεων είναι ο PCA (*Principal Component Analysis*).

Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (*Reinforcement Learning*) αποτελεί την κατηγορία της μηχανικής μάθησης που μιμείται πιο πιστά τον τρόπο μάθησης των ανθρώπων. Ένας πράκτορας ενισχυτικής μάθησης εκπαιδεύεται μέσω της αλληλεπίδρασης με το περιβάλλον του και επιλέγοντας δράσεις σε αυτό. Ο πράκτορας λαμβάνει θετική ανταμοιβή, όταν οι ενέργειες του οδηγούν σε επιθυμητά αποτελέσματα και αρνητική ανταμοιβή (τιμωρία), όταν οδηγούν σε ανεπιθύμητα αποτελέσματα. Μέσω δοκιμών και λαθών, ο πράκτορας μαθαίνει να παίρνει αποφάσεις που μεγιστοποιούν τις ανταμοιβές του. Η ενισχυτική μάθηση χρησιμοποιείται σε εφαρμογές όπως η ρομποτική, τα παιχνίδια και η διαχείριση πόρων. Μερικοί δημοφιλείς αλγόριθμοι ενισχυτικής μάθησης είναι ο Q-Learning, ο PPO (*Proximal Policy Optimization*) και ο SAC (*Soft Actor-Critic*).

Η ενισχυτική μάθηση αποτελεί την κατηγορία των αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται στην παρούσα εργασία και για αυτό, αναλύεται σε μεγαλύτερο βάθος στην επόμενη Ενότητα.

3.3 Ενισχυτική Μάθηση

3.3.1 Γενική επισκόπηση

Κεντρική ιδέα

Η Ενισχυτική Μάθηση (*Reinforcement Learning*) αποτελεί την κατηγορία της μηχανικής μάθησης, στην οποία ένας πράκτορας μαθαίνει από την εμπειρία του, καθώς αλληλεπιδράει με το περιβάλλον του. Ο πράκτορας ενισχυτικής μάθησης επιλέγει ελεύθερα ενέργειες και δέχεται ανάδραση για αυτές, υπό μορφή ανταμοιβής. Συγκεκριμένα, ο πράκτορας λαμβάνει θετική ανταμοιβή (επιβράβευση), όταν οι ενέργειες του οδηγούν σε επιθυμητά αποτελέσματα και αρνητική ανταμοιβή (τιμωρία), όταν οδηγούν σε ανεπιθύμητα αποτελέσματα. Έτσι, μέσω δοκιμών και λαθών (*trial and error*), ο πράκτορας μαθαίνει σταδιακά να παίρνει αποφάσεις που μεγιστοποιούν τις ανταμοιβές του. Η διαδικασία αυτή της εκπαίδευσης διαφέρει σημαντικά σε σχέση με τις δύο προηγούμενες κατηγορίες μηχανική μάθησης, αφού πλέον ο αλγόριθμος εκπαιδεύεται χωρίς κάποιο σταθερό σύνολο δεδομένων εισόδου. Αντίθετα, η εκπαίδευση στην ενισχυτική μάθηση είναι μία δυναμική διαδικασία, στην οποία ο πράκτορας εκτελεί ενέργειες και μεταβάλλει το περιβάλλον του. Επομένως, πρόκειται για μία ενεργή διαδικασία μάθησης, η οποία θυμίζει τον τρόπο με τον οποίο οι ζωντανοί οργανισμοί μαθαίνουν.

3.3 Ενισχυτική Μάθηση

Πράγματι, ας αναλογιστούμε το παράδειγμα της εκπαίδευσης ενός σκύλου. Στο σενάριο μας, ο ιδιοκτήτης του σκύλου επιθυμεί να του μάθει την ενέργεια «Κάτσε». Έτσι, ο ιδιοκτήτης του σκύλου δείχνει με το χέρι του το έδαφος και φωνάζει προς τον σκύλο «Κάτσε». Όσο ο σκύλος στέκεται όρθιος, ο ιδιοκτήτης του δίνει αρνητική ανταμοιβή (πχ του φωνάζει «κακό σκυλί», έχοντας τεντωμένο τον δείκτη του και δείχνοντας προς αυτό). Όταν ο σκύλος κάθεται, ο ιδιοκτήτης του δίνει θετική ανταμοιβή (πχ του χαιδεύει το κεφάλι ή του δίνει μία λιχουδιά). Έτσι, ο σκύλος μαθαίνει σταδιακά, πως όταν παρατηρεί τον άνθρωπο στην κατάσταση «Κάτσε» (δηλ. χέρι προς το έδαφος και προφορική εντολή), η ενέργεια του να καθίσει στο έδαφος του προσφέρει θετική ανταμοιβή και για αυτό την επιλέγει.

Γίνεται πλέον σαφές, πως στόχος της ενισχυτικής μάθησης είναι η εκπαίδευση του πράκτορα, ώστε να προβαίνει σε ενέργειες που μεγιστοποιούν την ανταμοιβή του. Μάλιστα, μία σημαντική παρατήρηση είναι πως ο πράκτορας πρέπει να μάθει να μεγιστοποιεί τη συνολική ανταμοιβή του, δηλ. να σκέπτεται μακροπρόθεσμα. Συγκεκριμένα, ενδέχεται μία ενέργεια του πράκτορα να οδηγήσει σε αρνητική ανταμοιβή όμεσα, όμως σε βάθος χρόνου να βοηθήσει τον πράκτορα να πετύχει μεγάλη θετική ανταμοιβή. Για παράδειγμα, στο παιχνίδι του σκακιού υπάρχει η στρατηγική της θυσίας (*sacrifice*), στην οποία ο παίκτης επιλέγει να χάσει ένα κομμάτι (πχ την βασίλισσα του), για να πετύχει στο μέλλον κάτι μεγαλύτερης αξίας (πχ ρουά-ματ στον αντίπαλο βασιλιά).

Πλεονεκτήματα

Η εκπαίδευση με ενισχυτική μάθηση έχει ορισμένα σημαντικά πλεονεκτήματα σε σχέση με τις πιο παραδοσιακές μεθόδους μηχανικής μάθησης, τα οποία περιγράφονται παρακάτω:

- Δεν χρειάζεται μεγάλα σύνολα δεδομένων: σε πολλά πεδία είναι δύσκολο να συλλεχθούν δεδομένα ή δεν υπάρχουν στον βαθμό που απαιτείται για μία αποδοτική εκπαίδευση επιβλεπόμενης μάθησης. Η ενισχυτική μάθηση αποτελεί μία εναλλακτική λύση, καθώς ο πράκτορας μαθαίνει μόνος του, χωρίς προηγούμενα δεδομένα.
- Δεν απαιτεί γνώση ειδικού στο πεδίο εφαρμογής: στην επιβλεπόμενη μάθηση, τα δεδομένα αποτελούν ζεύγη εισόδου-επιθυμητής εξόδου. Για παράδειγμα, για την εκπαίδευση ενός μοντέλου να παίζει σκάκι, τα δεδομένα θα ήταν της μορφής: κατάσταση σκακιέρας-κίνηση που έκανε ο παίκτης. Έτσι, το μοντέλο θα μάθαινε να παίζει σκάκι με τον τρόπο που παίζουν οι παίκτες στα δεδομένα εκπαίδευσης. Άρα, θα έπρεπε αυτοί οι παίκτες να είναι ειδικοί στο παιχνίδι, προκειμένου να πετύχει το μοντέλο υψηλή απόδοση. Αντίθετα, στην ενισχυτική μάθηση, οι σχεδιαστές ενός συστήματος αρκεί να έχουν βασική γνώση του πεδίου εφαρμογής του, για να εκπαιδεύσουν επιτυχώς έναν πράκτορα. Για παράδειγμα, στην περίπτωση του σκακιού, αρκεί οι σχεδιαστές να γνωρίζουν τους κανόνες του παιχνιδιού, ώστε πχ να επιβραβεύουν τον πράκτορα όταν κερδίζει κομμάτια του αντιπάλου και να τον τιμωρούν όταν χάνει κομμάτια του.

- προσφέρει **καινοτόμες λύσεις**: η επιβλεπόμενη μάθηση ουσιαστικά μιμείται τα δεδομένα εκπαίδευσης. Ναι μεν μπορεί να επιτυχεί υψηλότερη απόδοση από τον άνθρωπο (πχ σκάκι), όμως δεν μπορεί να μάθει μία εντελώς νέα προσέγγιση για την επίλυση του προβλήματος. Αντίθετα, οι αλγόριθμοι ενισχυτικής μάθησης μπορούν να προτείνουν εντελώς νέες και επαναστατικές λύσεις, τις οποίες δεν είχε σκεφτεί ποτέ κάποιος άνθρωπος. Ένα τέτοιο αξιοσημείωτο παράδειγμα συνέβη στη νική του AlphaGo, ενός συστήματος ενισχυτικής μάθησης έναντι του παγκόσμιου πρωταθλητή στο Go, Lee Sedol. Κατά τη διάρκεια του αγώνα, το AlphaGo έκανε μία ασυνήθιστη κίνηση (κίνηση 37) που αρχικά θεωρήθηκε λάθος από τους ειδικούς στο παιχνίδι. Μάλιστα, υπολογίστηκε ότι η πιθανότητα ένας άνθρωπος να προέβαινε σε αυτή την κίνηση, στη συγκεκριμένη θέση, ήταν ίση με 0.0001%. Ωστόσο, η κίνηση αποδείχθηκε εν τέλει δημιουργική και καθοριστική για τη νίκη του AlphaGo (Metz 2016).
- Είναι κατάλληλη για περιβάλλοντα **σειριακής λήψης αποφάσεων**: η ενισχυτική μάθηση μαθαίνει στον πράκτορα να μεγιστοποιεί τη συνολική ανταμοιβή σε βάθος χρόνου. Έτσι, είναι κατάλληλη για σενάρια όπου οι αποφάσεις είναι σειριακές και το αποτέλεσμα μίας ενέργειας επηρεάζει τις μελλοντικές αποφάσεις. Αντίθετα, στην επιβλεπόμενη μάθηση, το μοντέλο πραγματοποιεί ανεξάρτητες προβλέψεις σε κάθε βήμα, χωρίς να λαμβάνει υπόψη το πως αυτές θα επηρεάσουν τις μελλοντικές αποφάσεις.

Ωστόσο, η ενισχυτική μάθηση έχει και μειονεκτήματα, τα οποία δυσκολεύουν την επιτυχή εφαρμογή της. Πολλά από αυτά τα προβλήματα προέκυψαν και κατά τη διάρκεια της εκπαίδευσης του πράκτορα αυτόματης στάθμευσης. Για αυτό, περιγράφονται στο Κεφάλαιο —της εκπαίδευσης, στην Ενότητα ——.

Εφαρμογές

Η ενισχυτική μάθηση αποτελεί ουσιαστικά την επιστημή της λήψης αποφάσεων κι ως εκ τούτου, έχει εφαρμογές σε πολλά πεδία. Στο πεδίο των χρηματοοικονομικών, η εταιρία Goldman Sachs έχει ριζικά επενδυτικά στρατηγικά της (McMurray 2023). Στον τομέα της βιοτεχνολογίας, η εταιρία Atomwise χρησιμοποιεί την ενισχυτική μάθηση στην πλατφόρμα της AtomNet, που χρησιμοποιείται για την ανακάλυψη νέων φαρμάκων (Atomwise 2018). Στον χώρο της ρομποτικής, η εταιρία Boston Dynamics έχει ενσωματώσει την ενισχυτική μάθηση στα συστήματα ελέγχου του τετράποδου ρομπότ της, Spot και έχει βελτιώσει την αυτόνομη πλοιόγηση του σε πολύπλοκα περιβάλλοντα (BostonDynamics 2024).

Ωστόσο, το επικρατέστερο πεδίο εφαρμογής των αλγορίθμων ενισχυτικής μάθησης είναι τα παιχνίδια. Αυτό οφείλεται στο γεγονός ότι τα παιχνίδια αποτελούν έναν ασφαλές χώρο εκπαίδευσης πρακτόρων και ανάπτυξης αλγορίθμων, προτού αυτοί εφαρμοστούν σε προβλήματα του πραγματικού κόσμου. Μάλιστα, σε περιβάλλοντα προσομοιώσεων δεν υπάρχουν χρονικοί περιορισμοί, οπότε ο

3.3 Ενισχυτική Μάθηση

πράκτορας μπορεί να εκπαιδεύεται ασταμάτητα για μεγάλα χρονικά διαστήματα, κάτι που συχνά είναι απαραίτητο για την επιτυχία της ενισχυτικής μάθησης. Επίσης, τα παιχνίδια απαιτούν σε σημαντικό βαθμό νοητικές ικανότητες από τον παίκτη, κι έτσι αποτελούν μία καλή πλατφόρμα εκπαίδευσης αλγορίθμων τεχνητής νοημοσύνης.

Οι πρώτες απόπειρες έγιναν το 1959 από τον Arthur Samuel, ο οποίος ανέπτυξε ένα πρόγραμμα που έπαιζε το παιχνίδι της ντάμας (Samuel 1959). Στη συνέχεια, το 1992, ο Gerald Tesauro ανέπτυξε τον αλγόριθμο TD-Gammon με τη βοήθεια νευρωνικών δικτύων, ο οποίος έπαιζε τάβλι και κατάφερε να φτάσει σε επίπεδο ανάλογο των τριών κορυφαίων παικτών στον κόσμο (Tesauro 1994).

Ωστόσο, μέχρι και τη δεκαετία του 2010, υπήρχαν σημαντικοί περιορισμοί στην εφαρμογή της ενισχυτικής μάθησης σε πολύπλοκα προβλήματα ή προβλήματα μεγάλης διαστασιμότητας. Παρόλο που είχαν ήδη αναπτυχθεί αλγόριθμοι επίλυσης τέτοιων προβλημάτων, η μικρή διαθέσιμη υπολογιστική ισχύ της εποχής δεν επέτρεπε την εκπαίδευση των μοντέλων σε λογικά χρονικά διαστήματα. Το πρόβλημα αυτό υπερκεράστηκε από την τεχνολογική ανάπτυξη στον τομέα του υλικού (*hardware*) με χαρακτηριστικό παράδειγμα την παραλληλοποίηση των υπολογισμών στις μονάδες επεξεργασίας γραφικών (*GPU*), οι οποίες επιταχύνουν σημαντικά τη διαδικασία της εκπαίδευσης. Έτσι, ξεκίνησε η εποχή της βαθιάς ενισχυτικής μάθησης και άρχισαν να φαίνονται για πρώτη φορά οι πραγματικές δυνατότητες των τεχνητών νευρωνικών δικτύων. Ορισμένες επιτυχίες-ορόσημα της βαθιάς ενισχυτικής μάθησης σε περιβάλλοντα παιχνιδιών αναφέρονται με χρονολογική σειρά παρακάτω:

- 2012: η εταιρία DeepMind της Google ανέπτυξε το πρώτο σύγχρονο σύστημα βαθιάς ενισχυτικής μάθησης από, ο αλγόριθμος DQN (*Deep Q-Network*). Ο αλγόριθμος αυτός, εκπαιδεύτηκε ξεχωριστά στα 49 παιχνίδια της πλατφόρμας Atari2600, δεχόμενος ως είσοδο μόνο τα pixels της οθόνης και το σκορ του παιχνιδιού και κατάφερε να φτάσει σε επίπεδο συγκρίσιμο με αυτό ενός επαγγελματία δοκιμαστή παιχνιδιών (Mnih κ.ά. 2015).
- 2016: η εταιρία DeepMind παρουσίασε τον αλγόριθμο AlphaGo, για το παιχνίδι στρατηγικής Go (DeepMind 2016). Το Go, είναι ένα παιχνίδι πολύ πιο πολύπλοκο από το σκάκι, έχοντας σημαντικά μεγαλύτερο χώρο καταστάσεων κι έτσι, αποτελούσε πρόκληση για την τεχνητή νοημοσύνη. Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης δυσκολεύονταν να αξιολογήσουν όλες τις πιθανές κινήσεις, να αναπτύξουν ανθρώπινη δημιουργικότητα και συνολικά, να ανταγωνιστούν τους ανθρώπους. Όμως, ο αλγόριθμος AlphaGo κατάφερε να νικήσει σε μία σειρά παιχνιδιών τον θρυλικό παγκόσμιο πρωταθλητή στο Go, Lee Sedol. Αυτή η νίκη αποτέλεσε απόδειξη πως τα συστήματα βαθιάς ενισχυτικής μάθησης μπορούν να μάθουν να λύνουν τα πιο δύσκολα προβλήματα σε υψηλά περίπλοκα περιβάλλοντα.
- 2017: η εταιρία OpenAI ανέπτυξε μία παραλλαγή του αλγορίθμου Proximal Policy Optimization, τον αλγόριθμο OpenAI Five για την εκπαίδευση πρακτόρων στο παιχνίδι Dota2, ένα πολυπρακτορικό παιχνίδι (παίζεται από 2 ομάδες των 5 ατόμων), αβέβαιης πληροφορίας και με ιδιαίτερα πολύπλοκες καταστάσεις και ενέργειες. Οι πράκτορες του OpenAI Five

εκπαιδεύτηκαν για 10 μήνες και μέσω self-play, δηλαδή παίζοντας μεταξύ τους, και το 2019 κατάφεραν να νικήσουν τους τρέχοντες παγκόσμιους πρωταθλητές στο παιχνίδι (OpenAI 2019).

- **2020:** η εταιρία DeepMind παρουσίασε τον Agent57, μία βελτιωμένη έκδοση του αλγορίθμου DQN, ο οποίος χρησιμοποιεί έναν μετα-ελεγκτή για την προσαρμογή της εξερεύνησης και τη ρύθμιση της μακροπρόθεσμης έναντι της βραχυπρόθεσμης συμπεριφοράς του πράκτορα (Puigdomenech κ.ά. 2020). Ο Agent57 εκπαιδεύτηκε στα 57 παιχνίδια της πλατφόρμας Atari2600 και κατάφερε να ξεπεράσει τις επιδόσεις επαγγελματιών παικτών σε κάθε ένα από αυτά. Το σημαντικό σε αυτή την επιτυχία, είναι πως ήταν η πρώτη φορά που ένας αλγόριθμος κατάφερε κάτι ανάλογο στο σύνολο των 57 παιχνιδιών, τα οποία διακρίνονται για την πολυπλοκότητα και τη διαφορετικότητά τους.

3.3.2 Βασικές Έννοιες και Ορολογία

Έχοντας ήδη περιγράψει την γενικότερη ιδέα της ενισχυτικής μάθησης, στην ενότητα αυτή θα μελετήσουμε το πεδίο σε μεγαλύτερο βάθος, παρουσιάζοντας τις βασικές έννοιες και την ορολογία που χρησιμοποιείται. Αυτό θα βοηθήσει στην καλύτερη κατανόηση των βασικών αρχών της εκπαίδευσης πρακτόρων, τη μεθοδολογία που ακολουθείται, καθώς και τους στόχους αλλά και τα προβλήματα που προκύπτουν κατά την εκπαίδευση.

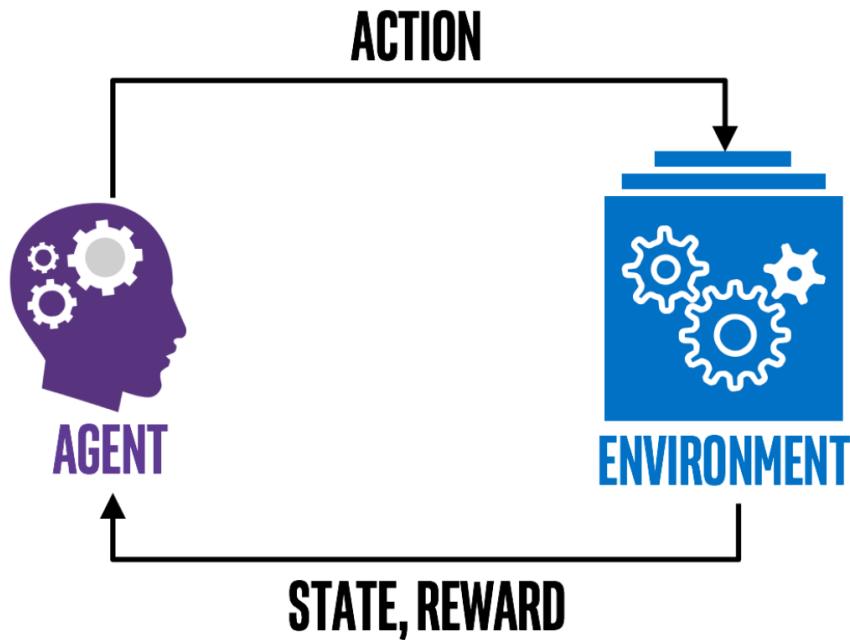
Κύκλος Ενισχυτικής Μάθησης

Ο κύκλος της ενισχυτικής μάθησης περιλαμβάνει την εξής διαδικασία: ο πράκτορας επιλέγει μία ενέργεια και την εκτελεί. Έτσι, μεταβάλει το περιβάλλον, δηλ. αυτό μεταβαίνει σε μία νέα κατάσταση. Έπειτα, ο πράκτορας δέχεται ως είσοδο τη νέα κατάσταση του περιβάλλοντος καθώς και την ανταμοιβή που προέκυψε από την ενέργειά του. Με βάση αυτές τις πληροφορίες, ο πράκτορας επιλέγει την επόμενη ενέργεια που θα εκτελέσει. Ο κύκλος αυτός φαίνεται και παραστατικά στην Εικόνα 3.4.

Μία επανάληψη της παραπάνω διαδικασίας ονομάζεται και βήμα (*step*) της εκπαίδευσης. Συνηθίζεται η εκπαίδευση ενός πράκτορα να χαρακτηρίζεται από το πλήθος των βημάτων στα οποία ο πράκτορας έχει εκπαιδευτεί. Στην πράξη, χρειάζονται μερικά εκατομμύρια βήματα εκπαίδευσης προκειμένου ο πράκτορας να φτάσει σε επιθυμητή απόδοση.

Επομένως, γίνεται πλέον κατανοητό πως κάθε πρόβλημα ενισχυτικής μάθησης αποτελείται από δύο βασικές οντότητες: το περιβάλλον και τον πράκτορα, καθώς επίσης και τρία κανάλια επικοινωνίας αυτών: της ανταμοιβής, των καταστάσεων και των ενεργειών. Ο όρος πράκτορας έχει ήδη αναλυθεί στην υποενότητα 3.1.2, ενώ οι υπόλοιποι όροι περιγράφονται λεπτομερώς στις επόμενες υποενότητες.

3.3 Ενισχυτική Μάθηση



Εικόνα 3.4. Κύκλος Ενισχυτικής Μάθησης (Lee 2019).

Περιβάλλον

Το περιβάλλον αποτελεί τον κόσμο, στον οποίο ο πράκτορας εκτελεί ενέργειες. Ένα περιβάλλον ενισχυτικής μάθησης πρέπει να ικανοποιεί την Μαρκοβιανή ιδιότητα. Προκειμένου να γίνει κατανοητή αυτή η ιδιότητα, πρέπει πρώτα να γίνει αναφορά στις Διαδικασίες Αποφάσεων Μαρκόβ.

Διαδικασίες Αποφάσεων Μαρκόβ

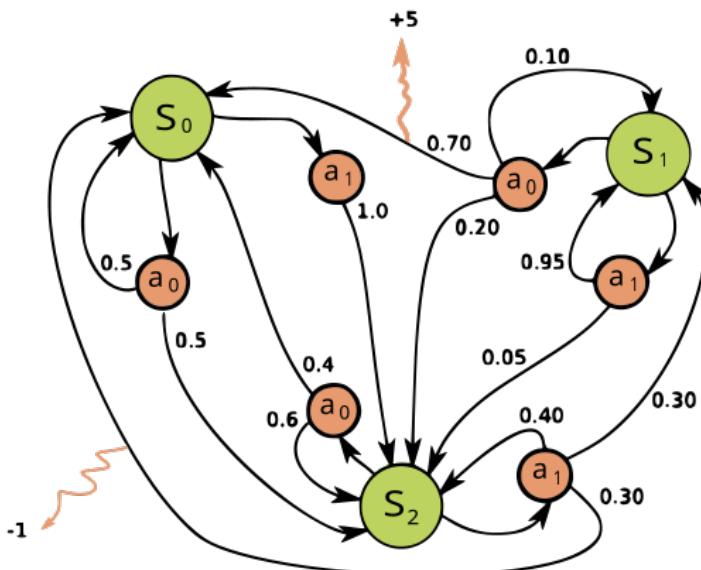
Οι διαδικασίες αποφάσεων Μαρκόβ (Markov Decision Processes - MDPs) αποτελούν ένα μαθηματικό πλαίσιο που χρησιμοποιείται για την περιγραφή ενός περιβάλλοντος σε προβλήματα επίτευξης ενός στόχου. Παρέχει μία μοντελοποίηση της λήψης αποφάσεων, σε καταστάσεις όπου τα αποτελέσματα είναι μερικώς τυχαία και μερικώς υπό τον έλεγχο του λήπτη αποφάσεων (δηλ. του πράκτορα). Αποτελούν επέκταση των αλυσίδων Μαρκόβ, με τη διαφορά ότι προσθέτουν ενέργειες και ανταμοιβές. Έτσι, τα MDPs χρησιμοποιούνται για να περιγράψουν ένα περιβάλλον στο οποίο θέλουμε να εφαρμόσουμε ενισχυτική μάθηση.

Περιγράφονται από την εξής διαδικασία: Ο πράκτορας αλληλεπιδρά με το περιβάλλον σε μία ακολουθία χρονικών στιγμών $t = 0, 1, 2, \dots$. Σε κάθε χρονική στιγμή t , το περιβάλλον βρίσκεται σε μία κατάσταση $s_t \in S$, όπου S είναι το σύνολο των καταστάσεων του περιβάλλοντος. Η κατάσταση αυτή δίνεται ως είσοδος στον πράκτορα, ο οποίος στη συνέχεια επιλέγει μία ενέργεια $a_t \in A$, όπου A είναι το σύνολο των ενεργειών που μπορεί να εκτελέσει. Το περιβάλλον ανταποκρίνεται στην ενέργεια του πράκτορα, μεταβαίνοντας σε μία νέα κατάσταση s_{t+1} και επιστρέφοντας στον

πράκτορα μία ανταμοιβή $r_{t+1} \in R$, όπου R είναι το σύνολο των πιθανών ανταμοιβών. Με τον τρόπο αυτό, μία διαδικασία απόφασης Μαρκόβ ορίζεται ως μία λίστα 4 στοιχείων (S, A, P, R) , όπου:

- S είναι το σύνολο των καταστάσεων του περιβάλλοντος (καλείται και χώρος καταστάσεων),
- A είναι το σύνολο των ενεργειών που μπορεί να εκτελέσει ο πράκτορας (καλείται και χώρος ενεργειών),
- P είναι η συνάρτηση πιθανότητας μετάβασης, όπου $P(s_{t+1}|s_t, a_t)$ είναι η πιθανότητα να μεταβεί το περιβάλλον στην κατάσταση s_{t+1} μετά την εκτέλεση της ενέργειας a_t στην κατάσταση s_t ,
- R είναι η συνάρτηση ανταμοιβής, ενώ το $R_{a_t}(s_t, s_{t+1})$ είναι η ανταμοιβή που λαμβάνει ο πράκτορας όταν μεταβεί από την κατάσταση s_t στην κατάσταση s_{t+1} μετά την εκτέλεση της ενέργειας a_t .

Ένα παράδειγμα μίας Διαδικασίας Απόφασης Μαρκόβ φαίνεται με τη χρήση ενός γράφου στην *Εικόνα 3.5*. Οι πράσινοι κόμβοι αντιστοιχούν στις καταστάσεις, οι πορτοκαλί κόμβοι στις ενέργειες, και τα βάρη των ακμών στις πιθανότητες μετάβασης. Επίσης, τα πορτοκαλί βέλη αντιστοιχούν στις ανταμοιβές που λαμβάνει ο πράκτορας μετά την εκτέλεση της ενέργειας.



Εικόνα 3.5. Παράδειγμα Διαδικασίας Απόφασης Μαρκόβ (Alvarez 2017)

Μαρκοβιανή Ιδιότητα

Η μαρκοβιανή ιδιότητα αναφέρεται στην ιδιότητα «αμνησίας» μίας στοχαστικής διαδικασίας, δηλαδή στο χαρακτηριστικό ότι η μελλοντική εξέλιξή της είναι ανεξάρτητη από το παρελθόν της.

3.3 Ενισχυτική Μάθηση

Συγκεκριμένα, δηλώνει ότι η επόμενη κατάσταση s_{t+1} εξαρτάται μόνο από την τρέχουσα κατάσταση s_t και την ενέργεια a_t που επιλέγει ο πράκτορας. Διοθέντος αυτών των δύο, είναι ανεξάρτητη από όλες τις προηγούμενες καταστάσεις και ενέργειες. Αυτό περιγράφεται μαθηματικά από την εξίσωση 3.1:

$$P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \dots, s_0, a_0) = P(s_{t+1}|s_t, a_t) \quad (3.1)$$

Η μαρκοβιανή ιδιότητα είναι κρίσιμη για την επιτυχή εφαρμογή της ενισχυτικής μάθησης, καθώς εξασφαλίζει ότι ο πράκτορας μπορεί να λάβει αποφάσεις με βάση μόνο την τρέχουσα κατάσταση του περιβάλλοντος, χωρίς να χρειάζεται να αποθηκεύσει όλες τις προηγούμενες καταστάσεις.

Πέραν όμως της μαρκοβιανής ιδιότητας, ένα περιβάλλον ενισχυτικής μάθησης περιγράφεται κι από άλλες ιδιότητες, οι οποίες οφείλονται στη διαφορετική φύση του κάθε προβλήματος και επηρεάζουν τη σημαντικά τη σχεδίαση του πράκτορα. Οι ιδιότητες αυτές είναι οι εξής:

Πλήρως ή μερικώς παρατηρήσιμο

Το περιβάλλον είναι πλήρως παρατηρήσιμο όταν ο πράκτορας έχει πρόσβαση στην πλήρη κατάστασή του κάθε χρονική στιγμή. Σε αντίθεση, όταν ο πράκτορας έχει πρόσβαση μόνο σε μέρος της κατάστασης του περιβάλλοντος, τότε το περιβάλλον είναι μερικώς παρατηρήσιμο. Ένα παράδειγμα ενός πλήρους παρατηρήσιμου περιβάλλοντος είναι το παιχνίδι του σκακιού, όπου οι παίκτες γνωρίζουν τη θέση όλων των πιονιών και των πιονιών του αντιπάλου. Αντίθετα, ένα παράδειγμα μερικώς παρατηρήσιμου περιβάλλοντος είναι το παιχνίδι του πόκερ, όπου οι παίκτες δεν γνωρίζουν τις κάρτες των αντιπάλων τους.

Αιτιοκρατικό ή στοχαστικό

Αιτιοκρατικό χαρακτηρίζεται το περιβάλλον στο οποίο η επόμενη κατάστασή του προσδιορίζεται με ακρίβεια από την τρέχουσα κατάστασή του και την ενέργεια του πράκτορα. Στην περίπτωση αιτιοκρατικού περιβάλλοντος, όταν ο πράκτορας βρίσκεται σε συγκεκριμένη κατάσταση και εκτελεί μία συγκεκριμένη ενέργεια, θα προκύπτει πάντα η ίδια ανταμοιβή και η ίδια επόμενη κατάσταση. Για παράδειγμα, παιχνίδια όπως το σκάκι είναι αιτιοκρατικά, καθώς η κίνηση ενός πιονιού σε συγκεκριμένη θέση θα οδηγήσει πάντα σε μία συγκεκριμένη κατάσταση του παιχνιδιού.

Όταν δεν υπάρχει αυτή η νομοτελειακή σχέση μεταξύ καταστάσεων και ενεργειών, αλλά υπάρχει και ένας βαθμός τυχαιότητας, το περιβάλλον είναι στοχαστικό. Τότε, ακόμα και όταν ο πράκτορας βρίσκεται σε συγκεκριμένη κατάσταση και εκτελεί μία συγκεκριμένη ενέργεια, δεν μπορεί να είναι βέβαιος για την ανταμοιβή που θα λάβει ή την επόμενη κατάσταση του περιβάλλοντος. Ένα παράδειγμα στοχαστικού περιβάλλοντος είναι το παιχνίδι του πόκερ, όπου οι κάρτες που θα αποκαλυφθούν στην επόμενη φάση του παιχνιδιού είναι τυχαίες.

Διακριτό ή συνεχές

Ένα περιβάλλον μπορεί να είναι διακριτό ή συνεχές ανάλογα με το πλήθος των δυνατών καταστάσεων

του. Όταν το πλήθος αυτό είναι πεπερασμένο, τότε το περιβάλλον είναι διακριτό. Για παράδειγμα, η τρίλιζα είναι ένα παιχνίδι με διακριτό περιβάλλον, καθώς οι δυνατές καταστάσεις του παιχνιδιού είναι πεπερασμένες και μάλιστα, λίγες.

Αντίθετα, όταν το πλήθος των δυνατών καταστάσεων είναι άπειρο, τότε το περιβάλλον είναι συνεχές. Ένα παράδειγμα συνεχούς περιβάλλοντος είναι το περιβάλλον της πλοιήγησης ενός ρομπότ, όπου η θέση, η ταχύτητα κι ο προσανατολισμός του ρομπότ μπορούν να λάβουν οποιαδήποτε τιμή σε ένα συνεχές πεδίο τιμών.

Μονοπρακτορικό ή πολυπρακτορικό

Ανάλογα με το πλήθος των πρακτόρων που συμμετέχουν στο περιβάλλον, αυτό χαρακτηρίζεται ως μονοπρακτορικό ή πολυπρακτορικό. Τα παζλ αποτελούν παραδείγματα μονοπρακτορικού περιβάλλοντος, ενώ αντίθετα αθλήματα όπως το ποδόσφαιρο είναι πολυπρακτορικά, καθώς συμμετέχουν πολλοί πράκτορες, οι οποίοι συνεργάζονται ή ανταγωνίζονται μεταξύ τους.

Επεισοδιακό ή ακολουθιακό

Το περιβάλλον μπορεί να είναι επεισοδιακό ή ακολουθιακό ανάλογα με τον τρόπο με τον οποίο ο πράκτορας αλληλεπιδρά με αυτό. Σε επεισοδιακά περιβάλλοντα, η αλληλεπίδραση μεταξύ πράκτορα χωρίζεται σε επεισόδια, τα οποία τερματίζονται μετά από την επίτευξη του επιθυμητού στόχου ή μετά από ένα συγκεκριμένο αριθμό βημάτων. Για παράδειγμα το σκάκι μπορεί να θεωρηθεί επεισοδιακό περιβάλλον, καθώς κάθε παιχνίδι αποτελεί ένα επεισόδιο, το οποίο τελειώνει είτε μέσω ρουά ματ ή όταν τελειώσει ο χρόνος ενός παίκτη.

Αντίθετα, σε ακολουθιακά περιβάλλοντα, η αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος δεν έχει φυσικό τέλος και ο πράκτορας συνεχίζει να αλληλεπιδρά με το περιβάλλον για αόριστο χρονικό διάστημα. Ένα παράδειγμα ακολουθιακού περιβάλλοντος είναι η πλοιήγηση ενός ρομπότ σε έναν άγνωστο χώρο.

Με αραίες ή πυκνές ανταμοιβές

Η διάκριση εδώ αφορά τον ρυθμό με τον οποίο ο πράκτορας λαμβάνει ανταμοιβές από το περιβάλλον. Σε περιβάλλοντα με αραίες ανταμοιβές, ο πράκτορας λαμβάνει ανταμοιβή σπάνια, μόνο όταν επιτύχει έναν συγκεκριμένο στόχο. Για παράδειγμα στην τρίλιζα, ο πράκτορας επιβραβεύεται μόνο όταν καταφέρνει να κερδίσει το παιχνίδι.

Αντίθετα, σε περιβάλλοντα με πυκνές ανταμοιβές, ο πράκτορας λαμβάνει ανταμοιβή πιο συχνά, προσφέροντας του έτσι πιο άμεση ανατροφοδότηση για την ποιότητα των ενεργειών του. Ένα παράδειγμα πυκνών ανταμοιβών είναι το κλασικό περιβάλλον εκπαίδευσης CartPole, οπου ο πράκτορας προσπαθεί να διατηρήσει σε ισορροπία σε ένα όρθιο κοντάρι. Στο περιβάλλον αυτό, όσο το κοντάρι παραμένει σε ισορροπία, ο πράκτορας λαμβάνει συνεχώς ανταμοιβές.

Ανταμοιβή

Σε κάθε ενέργεια του πράκτορα, το περιβάλλον του επιστρέφει μία τιμή, την ανταμοιβή. Πρόκειται για έναν αριθμό, ο οποίος μπορεί να είναι είτε αρνητικός (ο πράκτορας «τιμωρείται») ή θετικός (ο πράκτορας «επιβραβεύεται»). Ουσιαστικά, η ανταμοιβή δείχνει πόσο καλή ή κακή είναι η κατάσταση που βρίσκεται ο πράκτορας. Σκοπός του πράκτορα είναι να μεγιστοποιήσει την αθροιστική ανταμοιβή του σε βάθος χρόνου.

Η συνάρτηση η οποία υπολογίζει σε κάθε βήμα του πράκτορα την ανταμοιβή που λαμβάνει, ονομάζεται Συνάρτηση Ανταμοιβής (*Reward Function*). Η συνάρτηση ανταμοιβής μπορεί να είναι απλή, με τη λογική της επιβράβευσης ή τιμωρίας στο τέλος του παιχνιδιού, ή πιο πολύπλοκη, με περισσότερες επιβραβεύσεις και τιμωρίες, ώστε να οδηγήσει τον πράκτορα στην επιθυμητή συμπεριφορά. Για παράδειγμα, ας θεωρήσουμε το παιχνίδι της πλοϊγήσης σε έναν λαβύρινθο, όπου στόχος του πράκτορα είναι να βρει την έξοδο του. Μία απλή συνάρτηση ανταμοιβής θα μπορούσε να είναι η ανταμοιβή +1 όταν βρει την έξοδο και -1 εφόσον τελειώσει ο διαθέσιμος χρόνος του πράκτορα. Αντίθετα, μία πιο πολύπλοκη συνάρτηση ανταμοιβής θα μπορούσε να είναι η ανταμοιβή +1000 όταν βρει την έξοδο, +10 όταν πλησιάζει σε αυτήν, -10 όταν απομακρύνεται από αυτήν και -1000 εφόσον τελειώσει ο διαθέσιμος χρόνος του πράκτορα. Η δεύτερη, αναλυτικότερη προσέγγιση ονομάζεται Διαμόρφωση Ανταμοιβής (*Reward Shaping*) και συνήθως οδηγεί σε καλύτερα αποτελέσματα, καθώς παρέχει περισσότερη ανάδραση στον πράκτορα, καθοδηγώντας τον προς τον τελικό στόχο. Ωστόσο, χρειάζεται περισσότερη προσοχή από τον σχεδιαστή του συστήματος, προκειμένου να αποφευχθούν μη επιθυμητές συμπεριφορές του πράκτορα. Για αυτό, κάποιες φορές είναι προτιμότερο να διατηρηθεί απλή η συνάρτηση ανταμοιβής, ακόμα κι αν αυτό σημαίνει μεγαλύτερο χρόνο εκπαίδευσης.

Η ορθή σχεδίαση μίας συνάρτησης ανταμοιβής είναι είναι καθοριστική για την επιτυχία του πράκτορα στην εκπαίδευση. Ωστόσο, η διαδικασία αυτή αποδεικνύεται στην πράξη δύσκολη, καθώς περιέχει αρκετές προκλήσεις και απαιτεί προσεκτική ανάλυση του προβλήματος. Περισσότερες λεπτομέρειες σχετικά με τα προβλήματα που μπορεί να προκύψουν από τη σχεδίαση της συνάρτησης ανταμοιβής και τις πρακτικές που χρησιμοποιούνται για την αντιμετώπισή τους, δίνονται στο Κεφάλαιο 5, στις *Ενότητες — και — αντίστοιχα*.

Κατάσταση

Η κατάσταση (*state*) αποτελεί το σύνολο των πληροφοριών που δέχεται ο πράκτορας από το περιβάλλον, σε μία ορισμένη χρονική στιγμή. Ουσιαστικά, πρόκειται για την κωδικοποίηση της οπτικής αναπαράστασης του περιβάλλοντος, σε μορφή πληροφοριών που μπορούν να δοθούν ως είσοδο στον πράκτορα. Για παράδειγμα, όταν οι άνθρωποι παίζουν με ένα παιχνίδι, το οπτικό κανάλι είναι αυτό που λαμβάνει τις περισσότερες πληροφορίες για την κατάσταση του παιχνιδιού. Σε ένα

επιτραπέζιο παιχνίδι όπως η τρίλιζα, οι άνθρωποι βλέπουν την εικόνα του ταμπλό και με βάση αυτήν παίρνουν αποφάσεις. Όμως ένας πράκτορας χρειάζεται δεδομένα σε μορφή αριθμών ως είσοδο για το νευρωνικό του δίκτυο. Συνεπώς, η κατάσταση πρέπει να κωδικοποιηθεί σε ένα διάνυσμα αριθμών. Για την περίπτωση της τρίλιζας θα μπορούσαμε να είχαμε ένα διάνυσμα 9 θέσεων, όπου κάθε μία συμβολίζει ένα τετράγωνο του ταμπλό και μπορεί να πάρει 3 διακριτές τιμές: 1 αν στη θέση υπάρχει Ο, -1 αν υπάρχει X και 0 αν είναι άδεια. Έτσι, το διάνυσμα $[0, 0, 0, 0, -1, 0, 0, 0, 0]$ αντιστοιχεί σε ένα ταμπλό με X στη μεσαία θέση.

Στο παραπάνω παράδειγμα, το κάθε στοιχείο του διανύσματος κατάστασης μπορούσε να πάρει συγκεκριμένες, διακριτές τιμές. Έτσι, ο χώρος καταστάσεων που δημιουργείται είναι διακριτός (*Discrete state space*). Υπάρχουν όμως περιπτώσεις, στις οποίες το κάθε στοιχείο του διανύσματος κατάστασης μπορεί να πάρει οποιαδήποτε τιμή σε ένα συνεχές διάστημα. Σε αυτές τις περιπτώσεις, ο χώρος καταστάσεων είναι συνεχής (*Continuous state space*) και ο αριθμός των διαφορετικών καταστάσεων του πράκτορα είναι άπειρος. Για παράδειγμα, στην πλοήγηση ενός ρομπότ σε έναν άγνωστο χώρο, η κατάσταση του ρομπότ μπορεί να περιγραφεί από τη θέση του στο χώρο, την ταχύτητά του και τον προσανατολισμό του. Κάθε μία από αυτές τις παραμέτρους μπορεί να πάρει οποιαδήποτε τιμή σε ένα συνεχές διάστημα κι έτσι ο χώρος καταστάσεων είναι συνεχής. Ο χαρακτηρισμός του χώρου καταστάσεων ως διακριτός ή συνεχής είναι καθοριστικής σημασίας, διότι παίζει σημαντικό ρόλο στην επιλογή του αλγορίθμου εκπαίδευσης του πράκτορα.

Γενικά, είναι προτιμότερο οι πληροφορίες που δίνονται στον πράκτορα να περιορίζονται μόνο στις χρήσιμες σε αυτόν για την επίτευξη του στόχου του. Αυτό έχει ως αποτέλεσμα τη μείωση της διαστασιμότητας του προβλήματος, κάτι που επιταχύνει την εκπαίδευση του πράκτορα. Ωστόσο, η περιορισμένη πληροφορία μπορεί να οδηγήσει σε ανεπαρκή εκπαίδευση του πράκτορα, καθώς αυτός δεν έχει την πλήρη εικόνα του περιβάλλοντος και μπορεί να χάσει σημαντικές πληροφορίες. Επομένως, η σχεδίαση της κατάστασης πρέπει να γίνει με προσοχή, ώστε να εξασφαλιστεί η ισορροπία μεταξύ της πληροφορίας και της διαστασιμότητας του προβλήματος.

Ενέργεια

Ο όρος «ενέργεια» αντιπροσωπεύει μία πράξη που μπορεί να πραγματοποιήσει ο πράκτορας στο περιβάλλον του. Σε κάθε χρονική στιγμή t , ο πράκτορας πρέπει να επιλέξει μία ενέργεια από το σύνολο των διαθέσιμων ενεργειών του. Οι ενέργειες μπορεί να είναι διακριτές ή συνεχείς. Όπως και πριν, στο παιχνίδι της τρίλιζας, ο αριθμός των δυνατών κινήσεων του πράκτορα είναι πεπερασμένος και άρα ο χώρος ενεργειών είναι διακριτός (*Discrete action space*). Αντίθετα, στην πλοήγηση ενός ρομπότ σε έναν άγνωστο χώρο, δεν αρκεί το ρομπότ να επιλέξει να κινηθεί πχ γρήγορα προς τα δεξιά, αλλά απαιτείται μεγαλύτερη ακρίβεια στην κίνηση του. Έτσι, το ρομπότ επιλέγει συγκεκριμένη ταχύτητα και γωνία κίνησης, με αποτέλεσμα ο χώρος ενεργειών είναι συνεχής (*Continuous action space*). Η επιλογή των δυνατών ενεργειών του πράκτορα από τον σχεδιαστή του συστήματος είναι σημαντική,

3.3 Ενισχυτική Μάθηση

επειδή επηρεάζει την πολυπλοκότητα του προβλήματος και την επιλογή του αλγορίθμου εκπαίδευσης.

Πολιτική

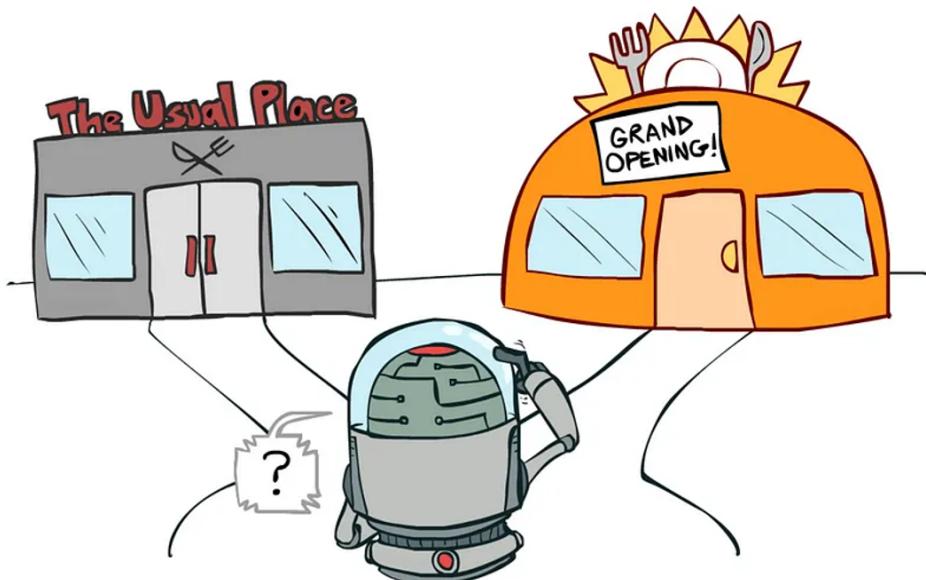
Η πολιτική του πράκτορα περιγράφει τον τρόπο με τον οποίο ο πράκτορας επιλέγει την επόμενη ενέργειά του σε κάθε κατάσταση. Από τεχνικής άποψης, η πολιτική του πράκτορα αποτελεί απλώς μία αντιστοιχίση της κάθε κατάστασης σε μία ενέργεια ή, αν θέλουμε να την παρομοιάσουμε με τον τρόπο που λειτουργούν οι άνθρωποι, πρόκειται για το σκεπτικό το οποίο χρησιμοποιεί ο πράκτορας για τη λήψη αποφάσεων. Η πολιτική αυτή μπορεί να είναι ντετερμινιστική ή στοχαστική. Στην πρώτη περίπτωση, συμβολίζεται ως $a_{t+1} = \pi(s_t)$, δηλ. η πολιτική αποτελεί την συνάρτηση που δέχεται ως είσοδο την κατάσταση του περιβάλλοντος και επιστρέφει την ενέργεια του πράκτορα. Τότε, όταν ο πράκτορας βρίσκεται στην ίδια κατάσταση, θα επιλέγει πάντα την ίδια ενέργεια. Στη δεύτερη περίπτωση, η πολιτική αντιστοιχίζει κάθε μία από τις δυνατές ενέργειες κάθε κατάστασης του πράκτορα, σε πιθανότητες. Συμβολίζεται ως $\pi(a_i|s_t)$ και αντιπροσωπεύει την πιθανότητα να επιλέξει την ενέργεια a_i ενώ βρίσκεται στην κατάσταση s_t . Εξυπακούεται ότι το άθροισμα των πιθανοτήτων όλων των πιθανών ενεργειών σε μία κατάσταση ισούται με τη μονάδα. Προκειμένου να γίνει καλύτερα κατανοητός ο τρόπος λειτουργίας μίας στοχαστικής πολιτικής, ας θεωρησούμε το παράδειγμα του κλασικού arcade παιχνιδιού «Snake» (φιδάκι). Στο παιχνίδι αυτό, ο πράκτορας έχει 4 δυνατές ενέργειες: πάνω, κάτω, αριστερά και δεξιά. Σε μία συγκεκριμένη κατάσταση, η στοχαστική πολιτική θα αντιστοιχίζει σε κάθε μία από τις 4 δυνατές κινήσεις του πράκτορα, μία πιθανότητα. Για παράδειγμα, αν οι πιθανότητες είναι: $P(\text{πάνω}) = 0.7$, $P(\text{κάτω}) = 0.2$, $P(\text{αριστερά}) = 0.05$ και $P(\text{δεξιά}) = 0.05$, τότε ο πράκτορας θα κινηθεί πάνω με πιθανότητα 0.7, κάτω με πιθανότητα 0.2 και αριστερά και δεξιά με πιθανότητα 0.05. Γενικά, οι στοχαστικές πολιτικές είναι πιο ευέλικτες από τις ντετερμινιστικές, καθώς επιτρέπουν στον πράκτορα να εξερευνήσει το περιβάλλον του και να ανακαλύψει νέες στρατηγικές. Ωστόσο, η επιλογή της στοχαστικής πολιτικής απαιτεί προσεκτική ανάλυση, καθώς μπορεί να οδηγήσει σε ανεπιθύμητες συμπεριφορές του πράκτορα.

Πλέον, γίνεται κατανοητό ότι η πολιτική που αναπτύσσει ο πράκτορας είναι ο παράγοντας που εν τέλει, καθορίζει την επιτυχία ή αποτυχία της εκπαίδευσης του. Το ζητούμενο είναι ο πράκτορας να αναπτύξει τη βέλτιστη πολιτική, η οποία θα τον οδηγήσει στην επίτευξη του στόχου του με τον πιο αποδοτικό τρόπο. Ωστόσο, η ανάπτυξη της βέλτιστης πολιτικής αποτελεί στην πράξη μία πρόκληση, καθώς απαιτεί την εξερεύνηση του περιβάλλοντος και την ανακάλυψη των στρατηγικών που θα οδηγήσουν στην επίτευξη του στόχου.

Έχοντας εξηγήσει την έννοια της πολιτικής, είμαστε πλέον σε θέση να παρουσιάσουμε το βασικότερο ίσως, πρόβλημα της ενισχυτικής μάθησης, το δίλημμα της εξερεύνησης έναντι της αξιοποίησης.

Δίλημμα Εξερεύνησης - Αξιοποίησης

Το δίλημμα της εξερεύνησης έναντι της αξιοποίησης (*exploration vs exploitation dilemma*) αποτελεί ένα από τα πιο δυσεπίλητα προβλήματα στο πεδίο της ενισχυτικής μάθησης. Το δίλημμα αναφέρεται στην ισορροπία μεταξύ της εξερεύνησης, δηλ. της ανακάλυψης νέων περιοχών και της εκμετάλλευσης, δηλ. της χρήσης της υπάρχουσας γνώσης. Έτσι, προκύπτει το ερώτημα: πρέπει ο πράκτορας να συνεχίσει να εφαρμόζει τις ενέργειες που γνωρίζει ότι λειτουργούν (αξιοποίηση) ή να δοκιμάσει νέες ενέργειες, προκειμένου να ανακαλύψει νέες στρατηγικές που ίσως είναι ακόμα πιο αποδοτικές (εξερεύνηση); Για παράδειγμα, πρέπει κάποιος να παραγγέλνει πάντα το ίδιο, γνωστό πιάτο στο αγαπημένο του εστιατόριο ή να δοκιμάσει κάτι καινούριο και διαφορετικό, με την ελπίδα να ανακαλύψει κάτι καλύτερο; Το παράδειγμα αυτό παρουσιάζεται και παραστατικά, στην Εικόνα 3.6.



Εικόνα 3.6. Το δίλημμα Εξερεύνησης - Αξιοποίησης (Parkinson 2019).

Προκειμένου να κατανοήσουμε σε βάθος το δίλημμα αυτό και τις επιπλοκές που έχει στην εκπαίδευση του πράκτορα, θα ξεκινήσουμε εξηγώντας τη διαδικασία που θα έπρεπε, ιδανικά, να ακολουθηθεί, ώστε ο πράκτορας να αναπτύξει την βέλτιστη πολιτική. Στη συνέχεια, θα αναδείξουμε τα προβλήματα που ανακύπτουν, τα οποία δυσκολεύουν την ομαλή εξέλιξη της παραπάνω διαδικασίας.

Στα πρώτα στάδια της εκπαίδευσης, επιθυμούμε ο πράκτορας να εξερευνήσει το περιβάλλον. Με τον όρο «εξερεύνηση», εννοούμε ο πράκτορας να δοκιμάσει τυχαίες ενέργειες σε διαφορετικές κατάστασεις, προκειμένου να αποκτήσει εμπειρία και να καταλάβει ποιές καταστάσεις είναι καλές και ποιές όχι. Δεν μας ενδιαφέρει ακόμα η απόδοση του πράκτορα, αλλά η ανάπτυξη μίας ισχυρής βάσης γνώσης. Είναι σημαντικό όλωστε να έχουμε κατά νου πως οι πράκτορες τεχνητής νοημοσύνης δεν έχουν όλες τις προηγούμενες γνώσεις κι εμπειρίες που έμεις, οι άνθρωποι, έχουμε αναπτύξει, ήδη

3.3 Ενισχυτική Μάθηση

από μικρή ηλικιά. Για παράδειγμα, στο περιβάλλον εκπαίδευσης αυτής της διπλωματικής εργασίας, στόχος του πράκτορα είναι η στάθμευση ενός αυτοκινητού σε μία ελεύθερη θέση. Εάν ο χειριστής του αυτοκινήτου ήταν άνθρωπος, αμέσως μόλις έβλεπε το περιβάλλον εκπαίδευσης (τον χώρο πάρκινγκ), θα είχε ήδη αρκετές χρήσιμες γνώσεις, όπως το πως κινείται ένα αυτοκίνητο, το ότι ο στόχος του είναι να παρκάρει το αυτοκίνητο στην ελεύθερη θέση ή το ότι η σύγκρουση με άλλα, σταθμευμένα αυτοκίνητα είναι -το λιγότερο- ανεπιθύμητη. Ωστόσο, ο πράκτορας δεν γνωρίζει ακόμα τίποτα από αυτά και πρέπει να τα ανακαλύψει μόνος του, εξερευνώντας το περιβάλλον και μαθαίνοντας από τις ανταμοιβές που πάίρνει. Έτσι, προς το παρόν, επιθυμούμε παραδείγματος χάριν, ο πράκτορας να συγκρούεται με άλλα αυτοκίνητα, ώστε να καταλάβει ότι αυτό δεν είναι επιθυμητό και να το αποφεύγει στα επόμενα στάδια της εκπαίδευσης του.

Μετά από ένα επαρκές διάστημα εξερεύνησης, ο πράκτορας πρέπει να περάσει στο στάδιο της αξιοποίησης. Στο στάδιο αυτό, ο πράκτορας δεν επιλέγει πια τυχαίες ενέργειες, αλλά εκμεταλλεύεται (*αξιοποιεῖ*) τη γνώση που έχει ήδη αποκτήσει, ώστε να επιλέγει σε κάθε κατάσταση την καλύτερη ενέργεια. Η μετάβαση από το ένα στάδιο στο άλλο πρέπει να γίνει σταδιακά, δηλ. επιθυμούμε ο πράκτορας να ελαττώσει βαθμιαία την τυχαιότητα των κινήσεων του, έως ότου τη μηδενίσει και βασίζεται εξ ολοκλήρου στην πολιτική του για τη λήψη αποφάσεων. Μόνο με αυτόν τον τρόπο, θα καταφέρει ο πράκτορας να ανακαλύψει τη βέλτιστη συμπεριφορά. Για παράδειγμα, στο περιβάλλον στάθμευσης αυτοκινήτων, επιθυμούμε ο πράκτορας να καταλάβει νωρίς στην εκπαίδευση, ότι το να πλησιάζει την ελεύθερη θέση είναι καλό. Έτσι, όσο βασίζεται όλο και περισσότερο στην πολιτική του για τη λήψη αποφάσεων, τόσο θα επιλέγει ενέργειες που θα τον κινούν πιο κοντά στην ελεύθερη θέση. Ταυτόχρονα ούμως, διατηρεί έναν βαθμό τυχαιότητας, ο οποίος αν και μειωμένος, τον αθείνει να εξακολουθεί να εξερευνεί νέες καταστάσεις και στρατηγικές. Έτσι, θα ανακαλύψει κάποτε τη μεγάλη επιβράβευση της στάθμευσης, θα εξερευνήσει διαφορετικούς τρόπους για να φτάνει σε αυτήν και στο τέλος της εκπαίδευσης, όταν θα αξιοποιεί αποκλειστικά την πολιτική του, θα επιλέγει πάντα τον βέλτιστο τρόπο στάθμευσης.

Με τη λογική της σταδιακής μείωσης της εξερεύνησης έναντι της αξιοποίησης, ο πράκτορας ωθείται να εξερευνήσει «προς τη σωστή κατεύθυνση». Αυτό σημαίνει πως δεν θέλουμε ο πράκτορας να σπαταλάει μεγάλο χρόνο εκπαίδευσης εξερευνώντας μη χρήσιμες καταστάσεις, όπως διαφορετικούς τρόπους να συγκρούεται με άλλα αυτοκίνητα, αλλά θέλουμε να εξερευνεί καταστάσεις χρήσιμες, που τον φέρνουν όλο πιο κοντά στον τελικό του στόχο.

Επομένως, για τη σωστή εκτέλεση της παραπάνω διαδικασίας, είναι κρίσιμη η ισορροπία μεταξύ της εξερεύνησης και της εκμετάλλευσης, δηλ. ο ρυθμός με τον οποίο ο πράκτορας μειώνει την τυχαιότητα των ενεργειών του. Εάν ο πράκτορας εξερευνήσει πολύ λίγο τον χώρο, τότε υπάρχει ο κίνδυνος να μην αναπτύξει καλή εικόνα του μοντέλου και όταν αρχίσει να ακολουθεί την πολιτική να οδηγείται σε λάθος κινήσεις. Αντίθετα, εάν ο χρόνος εξερεύνησης είναι πολύ μεγάλος, αυτό αυξάνει το κόστος και τον χρόνο εκτέλεσης του πειράματος. Επομένως, η επίτευξη της ισορροπίας αυτής αποτελεί μία από τις μεγαλύτερες προκλήσεις στον τομέα της ενισχυτικής μάθησης.

Είναι σημαντικό να σημειωθεί ότι η επιτυχία του πράκτορα στο στάδιο της αξιοποίησης εξαρτάται από την ποιότητα της εμπειρίας που έχει αποκτήσει κατά τη διάρκεια της εξερεύνησης. Εάν η εμπειρία αυτή είναι ανεπαρκής ή ανεπιθύμητη, τότε ο πράκτορας θα αντιμετωπίσει δυσκολίες στην επίτευξη του στόχου του.

Επομένως, για την επιτυχή εκτέλεση της ανωτέρω διαδικασίας, είναι κρίσιμη η ισορροπία μεταξύ εξερεύνησης και αξιοποίησης, δηλ. η ρύθμιση του ρυθμού μείωσης της τυχαιότητας των ενεργιών του πράκτορα. Η ισορροπία αυτή είναι στην πράξη πολύ δύσκολο να επιτευχθεί, καθώς δίνεται περισσότερη βαρύτητα στο ένα από τα δύο στάδια.

Συγκεκριμένα, εάν ο πράκτορας εξερευνήσει πολύ λίγο τον χώρο, τότε δεν θα αποκτήσει επαρκείς εμπειρίες και δεν θα ανακαλύψει τον βέλτιστο τρόπο για την επίτευξη του στόχου του - ή μπορεί και να μην ανακαλύψει καν τον στόχο του. Για να γίνει καλύτερα κατανοητό αυτό, ας επιστρέψουμε στο παράδειγμα της αυτόματης στάθμευσης κι ας εξετάσουμε μία τέτοια συμπεριφορά που συνέβη πολλές φορές στις εκπαίδευσις των πρακτόρων. Η επιλογή τυχαίων ενεργειών στα πρώτα στάδια της εκπαίδευσης, έχει ως αποτέλεσμα ο πράκτορας να συγκρούεται συχνά με άλλα αυτοκίνητα. Αν η εξερεύνηση σταματήσει πρόωρα, τώρα ο πράκτορας θα αρχίσει να εκμεταλλεύεται τη γνώση που έμαθε. Η γνώση αυτή, είναι μόνο ότι οι συγκρούσεις είναι ανεπιθύμητες, κι έτσι θα τις αποφεύγει. Ωστόσο, ο πράκτορας δεν πρόλαβε να μάθει πως η στάθμευση οδηγεί σε μεγάλη επιβράβευση κι έτσι, δεν επιχειρεί ποτέ να παρκάρει, αλλά κάνει απλώς κύκλους γύρω από την πίστα του παιχνιδιού. Τότε, λέμε πως ο πράκτορας έχει υιοθετήσει μία υποβέλτιστη πολιτική (*suboptimal policy*), ή έχει παγιδευτεί σε τοπικό μέγιστο (*local maximum*). Η δεύτερη έκφραση αναφέρεται στη συνάρτηση ανταμοιβής, καθώς γνωρίζουμε πως στόχος του πράκτορα είναι να τη μεγιστοποιεί, δηλ. να παίρνει ως ανταμοιβή το ολικό μέγιστο της. Όμως, όταν η εξερεύνηση του πράκτορα είναι ανεπαρκής, αυτός παίρνει ένα τοπικό μέγιστο της συνάρτησης ανταμοιβής, το οποίο είναι μικρότερο από το ολικό.

Από την άλλη, εάν ο χρόνος εξερεύνησης είναι πολύ μεγάλος, αυξάνεται ο χρόνος εκτέλεσης του πειράματος, επειδή ο πράκτορας αφιερώνει σημαντικό χρόνο σε μη χρήσιμες καταστάσεις. Ακόμα, υπάρχει ο κίνδυνος η απόδοση του πράκτορα να είναι ιδιαίτερα ασταθής, καθώς συνεχίζει να δοκιμάζει νέες ενέργειες αντί να βελτιώνει και να εξελίσσει τις γνωστές του στρατηγικές. Έτσι, μπορεί να μην καταφέρει να αναπτύξει μία σταθερή πολιτική, η οποία να τον οδηγεί στον στόχο του.

Μέχρι και σήμερα, δεν υπάρχει κάποια γενικά αποδεκτή λύση στο πρόβλημα της εξερεύνησης έναντι της αξιοποίησης. Έχουν προταθεί οριμένες τεχνικές για την επίτευξη της ζητούμενης ισορροπίας, όπως ο αλγόριθμος ϵ -greedy και η κανονικοποίηση της εντροπίας (*entropy regularization*), τις οποίες θα εξετάσουμε σε επόμενες ενότητες, όμως καμία δεν εγγυάται την επίτευξη της βέλτιστης πολιτικής.

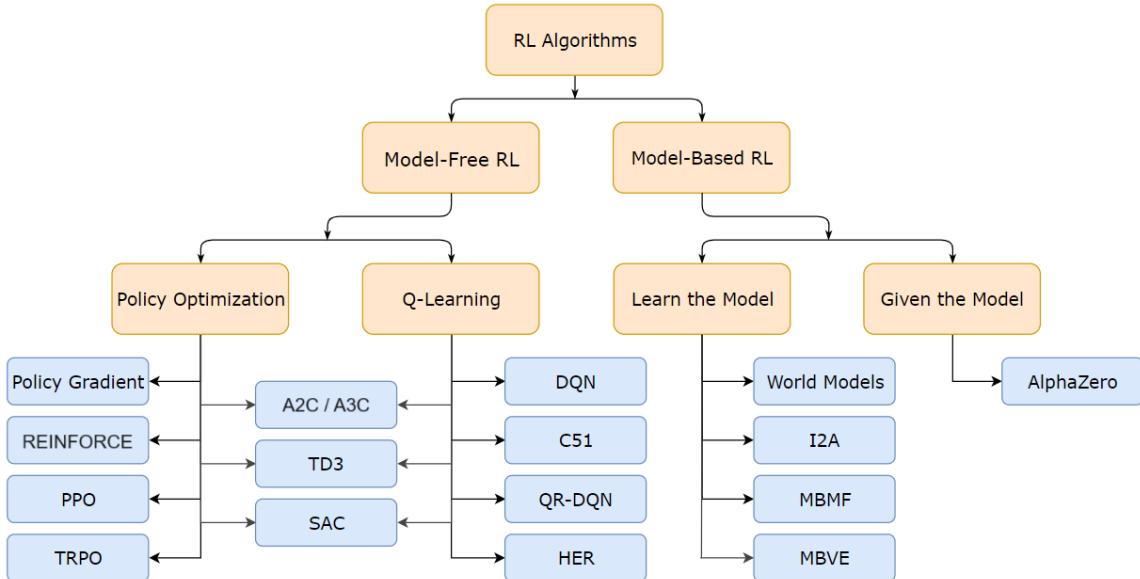
3.3.3 Κατηγορίες αλγορίθμων

Έχοντας κατανοήσει τις βασικές αρχές της ενισχυτικής μάθησης, μπορούμε πλέον να μελετήσουμε τους επικρατέστερους αλγορίθμους που χρησιμοποιούνται για την εκπαίδευση πρακτόρων. Στην

3.3 Ενισχυτική Μάθηση

Ενότητα αυτή, θα προβούμε σε μία επισκόπηση των διαφορετικών κατηγοριών αλγορίθμων και σε μεταγενέστερες ενότητες, θα εξετάσουμε τους αλγορίθμους που χρησιμοποιήθηκαν στα πλαίσια αυτής της εργασίας.

Μία συνοπτική ταξινόμηση των δημοφιλέστερων αλγορίθμων ενισχυτικής μάθησης παρουσιάζεται στην *Εικόνα 3.7*.



Εικόνα 3.7. Ταξινόμηση αλγορίθμων ενισχυτικής μάθησης (OpenAI 2018).

Παρατηρούμε πως μία πρώτη διάκριση των αλγορίθμων ενισχυτικής μάθησης γίνεται σε αλγορίθμους *Model-Free* και *Model-Based*.

Model-Free vs Model-Based

Η διάσπαση των αλγορίθμων στις δύο βασικές αυτές κατηγορίες εξαρτάται από την πρόσβαση ή μη του πράκτορα στο μοντέλο του περιβάλλοντος. Με τον όρο «μοντέλο του περιβάλλοντος» εννοούμε μία συνάρτηση που προβλέπει τις μεταβάσεις καταστάσεων και τις ανταμοιβές. Οι αλγόριθμοι model based απαιτούν την ύπαρξη ενός μοντέλου του περιβάλλοντος, ενώ οι model free όχι.

Model Based

Το μεγάλο πλεονέκτημα της γνώσης του μοντέλου έγκειται στο γεγονός ότι ο πράκτορας γνωρίζει όλες τις πιθανές ενέργειες του και τα αποτελέσματα αυτών. Επομένως, έχει τη δυνατότητα να σχεδιάσει την επόμενη κίνηση του, επιλέγοντας την καλύτερη από τις διαθέσιμες ενέργειες. Με τον τρόπο αυτό, οι αλγόριθμοι model based χρησιμοποιούν το μοντέλο για την ανάπτυξη μίας βέλτιστης πολιτικής. Όταν

αυτή η προσέγγιση λειτουργεί, μπορεί να οδηγήσει σε σημαντική βελτίωση της αποδοτικότητας σε σχέση με αλγορίθμους που δεν διαθέτουν μοντέλο.

Ωστόσο, το κύριο πρόβλημα αυτών των αλγορίθμων έγκειται στη δημιουργία του μοντέλου. Συγκεκριμένα, το μοντέλο αυτό είτε προυπάρχει και είναι διαθέσιμο στον πράκτορα, είτε πρέπει να δημιουργηθεί από τον ίδιο. Στην πλειοψηφία των προβλημάτων ενισχυτικής μάθησης, το μοντέλο του περιβάλλοντος είναι άγνωστο και πρέπει να αναπτυχθεί από τον πράκτορα κατά την εκπαίδευση του, μέσω των εμπειριών του. Αυτό παρουσιάζει αρκετές προκλήσεις, καθώς η πολυπλοκότητα του πραγματικού κόσμου καθιστά δύσκολη τη δημιουργία ενός ακριβού μοντέλου. Κάθε απόκλιση μεταξύ του μοντέλου και του πραγματικού κόσμου (*model bias*) μπορεί να οδηγήσει σε μειωμένη απόδοση του πράκτορα, κατά την εφαρμογή του στον πραγματικό κόσμο. Συνολικά, η εκμάθηση του μοντέλου είναι δύσκολη και μπορεί να αποτύχει, ακόμα και αν αφιερωθεί πολύς χρόνος και υπολογιστική ισχύς στην ανάπτυξή του.

Model Free

Οι αλγόριθμοι model free δεν απαιτούν γνώση ενός μοντέλου του περιβάλλοντος, αλλά στηρίζονται αποκλειστικά στην αλληλεπίδραση του πράκτορα με το περιβάλλον. Επομένως, ο πράκτορας χρησιμοποιεί την εμπειρία του για την ανάπτυξη της πολιτικής του. Το μειονέκτημα αυτής της προσέγγισης είναι ότι συνήθως απαιτείται περισσότερος χρόνος εκπαίδευσης, προεκιμένου ο πράκτορας να αναπτύξει μία αποδοτική πολιτική. Ωστόσο, η απουσία ανάπτυξης ενός μοντέλου του περιβάλλοντος, καθιστά τους αλγορίθμους model free πιο εύκολους στην υλοποίηση και τη ρύθμιση. Επιπλέον, οι αλγόριθμοι αυτοί αποδεικνύονται πιο σταθεροί και αξιόπιστοι, καθώς δεν επηρεάζονται από τυχόν σφάλματα και αποκλίσεις του μοντέλου. Έτσι, οι μέθοδοι χωρίς μοντέλο είναι πιο δημοφιλείς κι έχουν αναπτυχθεί και δοκιμαστεί περισσότερο από τις μεθόδους με μοντέλο.

Για τους λόγους που αναλύθηκαν παραπάνω, όλοι οι αλγόριθμοι που υλοποιήθηκαν στα πλαίσια αυτής της εργασίας ανήκουν στην κατηγορία των model free αλγορίθμων. Ως εκ τούτου, στη συνέχεια θα αναλύσουμε μόνο τις υποκατηγορίες των αλγορίθμων model free.

Κατηγορίες Model Free Αλγορίθμων

Υπάρχουν δύο βασικές προσεγγίσεις στους model free αλγορίθμους: οι αλγόριθμοι Εκτίμησης Αξίας (*Value Estimation*) και οι αλγόριθμοι Βελτιστοποίησης Πολιτικής (*Policy Optimization*). Υπάρχει όμως κι ένας τρίτος τύπος model free αλγορίθμων, οι αλγόριθμοι Δράστη-Κριτή (*Actor-Critic*), οι οποίοι συνδυάζουν στοιχεία από τις δύο προηγούμενες προσεγγίσεις. Κάθε ένα από αυτούς τους τύπους έχει τα δικά του πλεονεκτήματα και μειονεκτήματα, και είναι κατάλληλος για διαφορετικούς τύπους προβλημάτων. Οι 3 αυτοί τύποι φαίνονται στο κάτω αριστερά τμήμα της Εικόνας 3.7¹ και εξετάζονται στη συνέχεια.

¹Οι αλγόριθμοι Εκτίμησης Αξίας παρουσιάζονται στην εικόνα ως Q-Learning αλγόριθμοι, ενώ οι αλγόριθμοι που βρίσκονται μεταξύ των δύο βασικών προσεγγίσεων αποτελούν τους αλγορίθμους Δράστη-Κριτή.

Αλγόριθμοι Εκτίμησης Αξίας

Οι αλγόριθμοι Εκτίμησης Αξίας (*Value Estimation*) χωρίζονται σε δύο υποκατηγορίες, τους αλγορίθμους Εκτίμησης Αξίας Κατάστασης (*State Value Estimation*) και τους αλγορίθμους Εκτίμησης Αξίας Ζεύγους Κατάστασης-Ενέργειας (*State-Action Value Estimation*). Η λογική πίσω από αυτούς τους αλγορίθμους όμως παραμένει η ίδια και στις 2 περιπτώσεις και αναλύεται στη συνέχεια.

Οι αλγόριθμοι Εκτίμησης Αξίας Καταστάσεων επικεντρώνονται στην εκμάθηση της αξίας των καταστάσεων. Η βασική ιδέα είναι η ανάπτυξη μίας συνάρτησης αξίας (*value function*), η οποία συμβολίζεται ως $V_\pi(s)$ και θα εκτιμάει την ποιότητα μίας κατάστασης. Ο όρος «αξία» ή «ποιότητα» μίας κατάστασης αναφέρεται στην αναμενόμενη αθροιστική ανταμοιβή που θα λάβει ο πράκτορας, ξεκινώντας από την κατάσταση αυτή και ακολουθώντας έπειτα την πολιτική που έχει αναπτύξει.

Κατά παρόμοιο τρόπο, οι αλγόριθμοι Εκτίμησης Αξίας Ζευγών Κατάστασης-Ενέργειας εστιάζουν στην εκμάθηση της αξίας των ζευγών κατάστασης-ενέργειας. Η συνάρτηση αξίας σε αυτή την περίπτωση συμβολίζεται ως $Q_\theta(s, a)$ και εκτιμά την ποιότητα της ενέργειας a στην κατάσταση s , δηλ. την αναμενόμενη αθροιστική ανταμοιβή που θα λάβει ο πράκτορας, εάν επιλέξει την ενέργεια a στην κατάσταση s και ακολουθήσει έπειτα την πολιτική που έχει αναπτύξει. Οι ενημερώσεις της συνάρτησης αξίας γίνονται *off-policy*, το οποίο σημαίνει ότι κάθε ενημέρωση μπορεί να χρησιμοποιήσει δεδομένα που συλλέχθηκαν σε οποιοδήποτε σημείο της εκπαίδευσης, ανεξάρτητα από το πώς εξερευνούσε τότε ο πράκτορας το περιβάλλον. Χαρακτηριστικό παράδειγμα αυτής της προσέγγισης είναι ο αλγόριθμος *Q-Learning* και για αυτό, οι συγκεκριμένοι αλγόριθμοι συχνά αναφέρονται και ως *Q-Learning* αλγόριθμοι.

Τα μεγαλύτερο πλεονέκτημα των αλγορίθμων Εκτίμησης Αξίας είναι η απλότητά τους. Ακόμα, είναι ιδιαίτερα αποτελεσματικοί σε περιβάλλοντα με διακριτούς χώρους καταστάσεων και ενεργειών. Στα περιβάλλοντα αυτά, είναι σημαντικά πιο αποδοτικοί στον χρόνο εκπαίδευσης σε σχέση με τις άλλες κατηγορίες αλγορίθμων, επειδή μπορούν να επαναχρησιμοποιήσουν τα ίδια δεδομένα πολλές φορές. Όμως, οι αλγόριθμοι Εκτίμησης Αξίας βελτιστοποιούν μόνο έμμεσα την απόδοση του πράκτορα, μέσω της ενημέρωσης της συνάρτησης αξίας. Επίσης, παρουσιάζουν δυσκολία στην αντιμετώπιση μεγάλων ή συνεχών χώρων καταστάσεων και ενεργειών, αφού απαιτούν την αποθήκευση μίας τιμής για κάθε ζεύγος κατάστασης-ενέργειας. Έτσι, συνολικά, τείνουν να είναι λιγότερο σταθεροί από τους αλγορίθμους των άλλων δύο κατηγοριών.

Αλγόριθμοι Βελτιστοποίησης Πολιτικής

Οι αλγόριθμοι Βελτιστοποίησης Πολιτικής (*Policy Optimization*) μαθαίνουν απευθείας μία πολιτική π_θ , που μεγιστοποιεί την αναμενόμενη αθροιστική ανταμοιβή, χωρίς να απαιτούν την ανάπτυξη συνάρτησης αξιών. Αντίθετα, επικεντρώνονται στη ρύθμιση των παραμέτρων θ της πολιτικής, προκειμένου να βελτιστοποίησουν την απόδοση τους. Αυτή η ρύθμιση πραγματοποιείται *on-policy*, δηλαδή κάθε ενημέρωση χρησιμοποιεί μόνο δεδομένα που συλλέχθηκαν ενώ ο πράκτορας ενεργούσε σύμφωνα με την πιο πρόσφατη έκδοση της πολιτικής του. Όπως αναφέρθηκε και νωρίτερα, η

πολιτική μπορεί να είναι ντετερμινιστική ή στοχαστική. Μία ντετερμινιστική πολιτική συμβολίζεται ως $a_{t+1} = \pi_\theta(s_t)$ και υποδηλώνει ότι σε συγκεκριμένη κατάσταση, θα επιλέγεται κάθε φορά η ίδια ενέργεια. Αντίθετα, μία στοχαστική πολιτική συμβολίζεται ως $\pi_\theta(a_i|s_t)$ και εισάγει ένα βαθμό τυχαιότητας στη λήψη αποφάσεων από τον πράκτορα, αφού η επιλεγμένη ενέργεια προκύπτει από μία κατανομή πιθανοτήτων. Παραδείγματα αλγορίθμων αυτής της κατηγορίας αποτελούν οι κλασικοί αλγόριθμοι REINFORCE και PPO.

Το μεγαλύτερο πλεονέκτημα που έχουν οι αλγόριθμοι της κατηγορίας βελτιστοποίησης πολιτικής είναι ότι βελτιστοποιούν άμεσα την πολιτική τους, δηλ. τον στόχο της εκπαίδευσης. Αυτό το γεγονός τους καθιστά σταθερούς και αξιόπιστους. Ακόμα, αξιοσημείωτη είναι η απόδοση τους σε περιβάλλοντα με συνεχείς χώρους καταστάσεων ή και ενεργειών. Ωστόσο, συχνά απαιτούν μεγάλο χρόνο εκπαίδευσης για την επίτευξη σταθερής μάθησης. Επιπλέον, απαιτούν πολύ προσεκτική ρύθμιση των παραμέτρων τους, επειδή είναι επιρρεπείς σε τοπικά ελάχιστα.

Αλγόριθμοι Δράστη - Κριτή

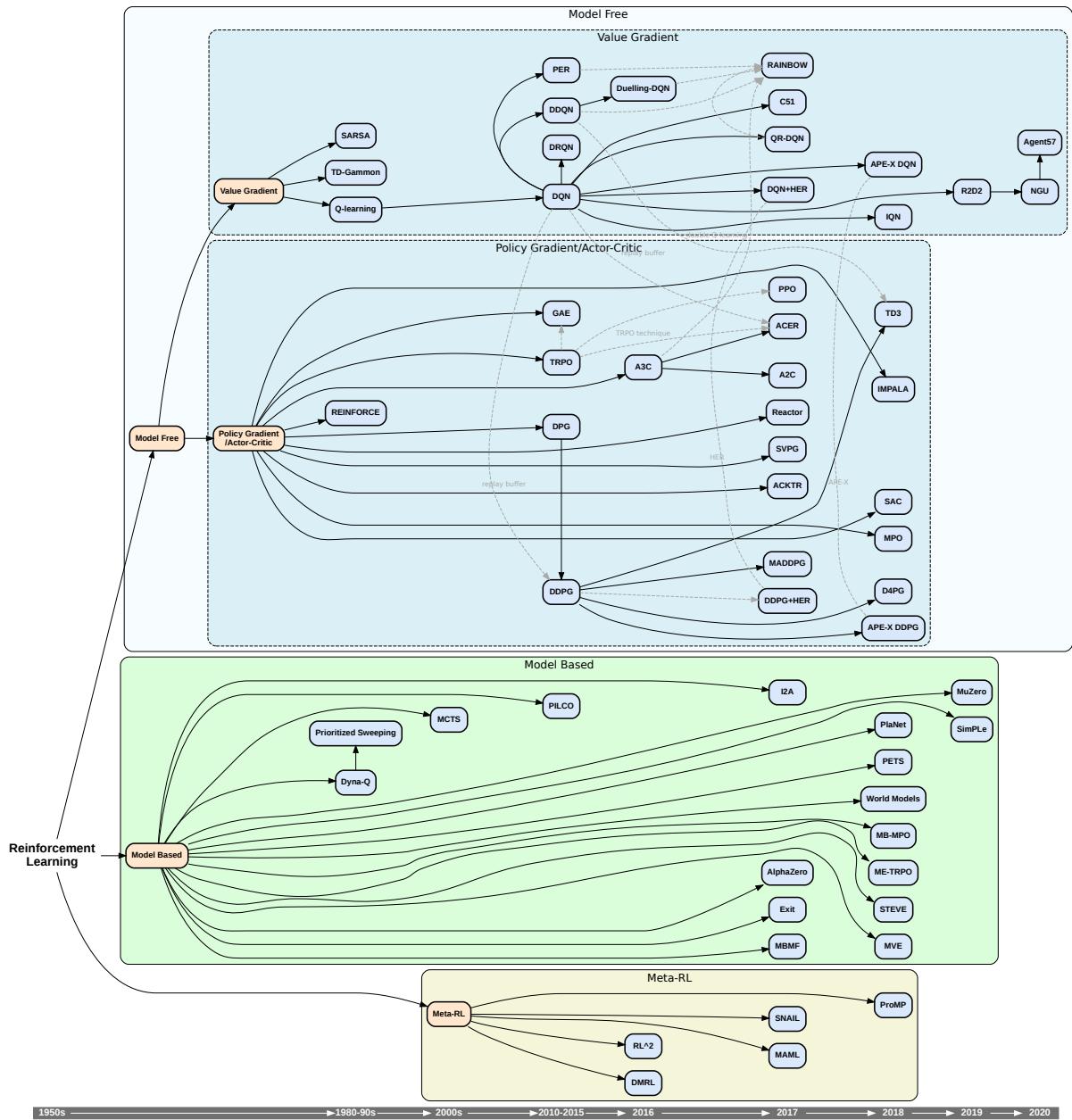
Οι αλγόριθμοι Δράστη-Κριτή (*Actor-Critic*) συνδυάζουν στοιχεία από τις δύο προηγούμενες κατηγορίες. Έτσι, προσπαθούν να εκμεταλλευτούν τα πλεονεκτήματα αλλά και να αποφύγουν τις αδυναμίες της κάθε προσέγγισης. Αποτελούνται από δύο ξεχωριστά νευρωνικά δίκτυα: ένα δίκτυο-δράστη και ένα δίκτυο-κριτή. Ο δράστης αποφασίζει ποια ενέργεια να πάρει ο πράκτορας και έπειτα ο κριτής αξιολογεί την επιλεγμένη ενέργεια με την εκτίμηση της συνάρτησης αξίας. Στη συνέχεια, ο κριτής παρέχει στον δράστη ανατροφοδότηση για να βελτιώσει την πολιτική του. Μάλιστα, ο κριτής βοηθά στη μείωση της διακύμανσης στις ενημερώσεις της πολιτικής, οδηγώντας σε πιο σταθερή μάθηση. Παραδείγματα αλγορίθμων αυτής της κατηγορίας αποτελούν οι SAC και TD3.

Το μεγαλύτερο πλεονέκτημα των αλγορίθμων Δράστη-Κριτή είναι πως αξιοποιούν τόσο τις τεχνικές της εκτίμησης αξίας, όσο και της βελτιστοποίησης πολιτικής. Έτσι, η διαδικασία αυτή επιτυγχάνει συνήθως μία καλή ισορροπία μεταξύ εξερεύνησης και αξιοποίησης, οδηγώντας σε πιο σταθερή και αποτελεσματική μάθηση. Χρησιμοποιείται ευρύτερα σε προβλήματα που περιλαμβάνουν πολύπλοκα περιβάλλοντα και μεγάλους χώρους καταστάσεων. Ωστόσο, οι αλγόριθμοι αυτοί είναι πιο πολύπλοκοι στην υλοποίηση και τη ρύθμιση, λόγω της αλληλεπίδρασης μεταξύ του δράστη και του κριτή. Μπορούν ακόμα να υποφέρουν από αστάθεια ή και απόκλιση, εάν ο κριτής παρέχει κακές εκτιμήσεις.

Τέλος, αξίζει να γίνει ειδική μνεία σε μία άλλη, λιγότερο γνωστή απόπειρα ταξινομήσης του μεγάλου πλήθους των αλγορίθμων ενισχυτικής μάθησης από τον (Prijono 2020), η οποία παρουσιάζεται στην Εικόνα 3.8.

Στην εικόνα αυτή, τα βέλη που συνδέουν δύο αλγορίθμους υποδηλώνουν ότι ο ένας αλγόριθμος αποτελεί βελτίωση του άλλου. Συγκεκριμένα, οι συνεχείς γραμμές υποδηλώνουν ισχυρή σύνδεση

3.3 Ενισχυτική Μάθηση



Εικόνα 3.8. Εκτενέστερη ταξινόμηση αλγορίθμων ενισχυτικής μάθησης (Prijono 2020).

μεταξύ των δύο αλγορίθμων, ενώ οι διακεκομμένες γραμμές υποδηλώνουν πιο ασθενή σύνδεση. Έτσι, καταλαβαίνουμε για παράδειγμα, πως οι αλγόριθμοι PPO, TD3 που χρησιμοποιηθήκαν στην παρούσα εργασία, αποτελούν βελτιώσεις των αλγορίθμων TRPO και DDPG αντίστοιχα. Επιπλέον, στο κάτω μέρος της εικόνας παρουσιάζεται το χρονολόγιο της δημοσίευσης των αλγορίθμων, προκειμένου να γίνει κατανοητή η εξέλιξη των αλγορίθμων στο χρόνο. Μάλιστα, στον σύνδεσμο που υπάρχει στη βιβλιογραφία, υπάρχει ένα πλήρες αποθετήριο, όπου για κάθε αλγόριθμο μπορεί κάνεις να βρει την αντίστοιχη δημοσίευση, καθώς και άλλες χρήσιμες πληροφορίες. Για αυτό, προτρέπω τους ενδιαφερόμενους αναγνώστες να επισκεφτούν και να χρησιμοποιήσουν το συγκεκριμένο αποθετήριο, για να περιηγηθούν γρήγορα και εύκολα στον χώρο των αλγορίθμων ενισχυτικής μάθησης.

3.3.4 Ο αλγόριθμος *Q-learning*

Ο αλγόριθμος *Q-Learning* προτάθηκε από τον Chris Watkins το 1989, στη διδακτορική του διατριβή (Watkins 1989). Είναι ένας από τους πιο διαδεδομένους αλγορίθμους ενισχυτικής μάθησης και αποτέλεσε τη βάση για πολλές εξελίξεις στον τομέα. Ανήκει στην κατηγορία των model free αλγορίθμων και πιο συγκεκριμένα, στην κατηγορία των αλγορίθμων εκτίμησης αξίας ζευγών κατάστασης-ενέργειας.

Μεθοδολογία αλγορίθμου

Όπως κι οι υπόλοιποι αλγόριθμοι της κατηγορίας αξίας ζευγών κατάστασης-ενέργειας, ο *Q-Learning* επικεντρώνεται στην εκμάθηση της αξίας των ζευγών κατάστασης-ενέργειας. Η συνάρτηση αξίας που αναπτύσσει συμβολίζεται ως $Q(s, a)$, όπου το «*Q*» αντιπροσωπεύει την ποιότητα (*Quality*) της ενέργειας a στην κατάσταση s , δηλ. πόσο χρήσιμη είναι η συγκεκριμένη ενέργεια στη συγκεκριμένη κατάσταση, στο να μεγιστοποιήσει τις μελλοντικές ανταμοιβές. Συνήθως, χρησιμοποιείται ένας πίνακας (*Q Table*) για την αποθήκευση των τιμών Q της συνάρτησης $Q(s, a)$, όπου κάθε γραμμή αντιστοιχεί σε μία κατάσταση και κάθε στήλη σε μία ενέργεια, οπως φαίνεται στην Εικόνα 3.9.

Το πιο σημαντικό κομμάτι του αλγορίθμου *Q-Learning* είναι η κατασκευή της συνάρτησης $Q(s, a)$. Η κατασκευή αυτή γίνεται μέσω της επαναληπτικής ενημέρωσης των τιμών Q .

Αρχικά, ας δούμε πως προκύπτει μία τιμή Q για ένα ζεύγος κατάστασης-ενέργειας. Η τιμή Q για το ζεύγος (s, a) υπολογίζεται κατά τη διάρκεια της εκπαίδευσης από την εξίσωση 3.2 (εξίσωση *Bellman*):

$$Q(s, a) = R(s, a) + \gamma \cdot \max_a Q(s', a) \quad (3.2)$$

Η εξίσωση αυτή, δηλώνει ότι η αξία του ζεύγους (s, a) είναι ίση με την ανταμοιβή που προκύπτει από την εκτέλεση της ενέργειας a στην κατάσταση s (συμβολίζεται ως $R(s, a)$), συν την αναμενόμενη ανταμοιβή που προκύπτει εκτελώντας την καλύτερη ενέργεια a σε όλες τις επόμενες καταστάσεις

3.3 Ενισχυτική Μάθηση

| | Actions | | | |
|----------------|-------------------------------------|-------------------------------------|-----|-------------------------------------|
| | A ₁ | A ₂ | ... | A _M |
| S ₁ | Q(S ₁ , A ₁) | Q(S ₁ , A ₂) | | Q(S ₁ , A _M) |
| S ₂ | Q(S ₂ , A ₁) | Q(S ₂ , A ₂) | | Q(S ₂ , A _M) |
| : | | | .. | : |
| S _N | Q(S _N , A ₁) | Q(S _N , A ₂) | ... | Q(S _N , A _M) |

Εικόνα 3.9. Πίνακας Q για την αποθήκευση των τιμών της συνάρτησης $Q(s, a)$ (Baeldung 2023).

s' (συμβολίζεται ως $\max_a Q(s', a)$), πολλαπλασιασμένη επί τον συντελεστή γ . Ο συντελεστής αυτός ονομάζεται παράγοντας έκπτωσης (*discount factor*), παίρνει τιμές στο διάστημα $[0, 1]$ και χρησιμοποιείται για τον καθορισμό της σημασίας των μελλοντικών ανταμοιβών. Συγκεκριμένα, οι μελλοντικές ανταμοιβές είναι λιγότερο πολύτιμες από τις τρέχουσες ανταμοιβές, και για αυτό ο παράγοντας γ τις μετριάζει. Όταν παίρνει τιμές κοντά στο 0, τότε ο πράκτορας συμπεριφέρεται πιο κοντόφθαλμα, ενώ όσο η τιμή του παράγοντα πλησιάζει το 1, τόσο ο πράκτορας προσπαθεί να μεγιστοποιήσει το μακροπρόθεσμο κέρδος του.

Ωστόσο, όπως αναφέραμε, η ενημέρωση των τιμών Q γίνεται επαναληπτικά κατά τη διάρκεια της εκπαίδευσης. Έτσι, στην αρχή της εκπαίδευσης, οι τιμές Q όλων των ζευγών κατάστασης-ενέργειας αρχικοποιούνται σε μία τιμή (συνήθως 0). Έπειτα, όταν ο πράκτορας επισκέπτεται μία κατάσταση s , εκτελεί μία ενέργεια a και λαμβάνει μία ανταμοιβή $R(s, a)$. Τότε, η τιμή Q του ζεύγους (s, a) ενημερώνεται σύμφωνα με την εξίσωση 3.3:

$$Q^{new}(s, a) = (1 - \alpha) \cdot Q^{old}(s, a) + \alpha \left[R(s, a) + \gamma \cdot \max_a Q(s', a) \right] \quad (3.3)$$

Επειδή αυτή η εξίσωση χρησιμοποιεί τη διαφορά χρησιμότητας μεταξύ διαδοχικών καταστάσεων (κι επομένως, διαδοχικών χρόνων), ονομάζεται και Εξίσωση Ενημέρωσης Χρονικών Διαφορών (*Temporal*

Difference Update Rule). Παρατηρούμε ότι η εξίσωση TD Update Rule χρησιμοποιεί την εξίσωση Bellman. Συγκεκριμένα, η νέα τιμή Q του ζεύγους (s, a) συμβολίζεται ως $Q^{new}(s, a)$ και προκύπτει από:

- τον όρο $(1 - \alpha) \cdot Q^{old}(s, a)$, που αντιπροσωπεύει την παλιά τιμή Q του ζεύγους (s, a) , δηλ. αυτήν που ήδη υπήρχε στον πίνακα Q , πολλαπλασιασμένη με τον συντελεστή $(1 - \alpha)$ και
- τον όρο $\alpha [R(s, a) + \gamma \cdot \max_a Q(s', a)]$, που αντιπροσωπεύει την τιμή Q του ζεύγους (s, a) που μόλις υπολογίστηκε από την εξίσωση Bellman, πολλαπλασιασμένη με τον συντελεστή α .

Ο συντελεστής α ονομάζεται ρυθμός μάθησης (*learning rate*) και καθορίζει σε ποιό βαθμό, οι νεότερες, πιο πρόσφατες πληροφορίες αντικαθιστούν τις παλιές, δηλ. καθορίζει τον ρυθμό με τον οποίο ο πράκτορας μαθαίνει. Ο ρυθμός μάθησης παίρνει κι αυτός τιμές στο διάστημα $[0, 1]$ και αποτελεί μία ακόμα κρίσιμη υπερπαράμετρο που πρέπει να ρυθμιστεί με προσοχή, καθώς επηρεάζει σημαντικά τη σύγκλιση του αλγορίθμου. Εάν πάρει την τιμή 0, τότε ο πίνακας Q δεν ενημερώνεται καθόλου (δηλ. ο πράκτορας δεν μαθαίνει τίποτα καινούργιο), ενώ εάν πάρει την τιμή 1, τότε ο πίνακας Q ενημερώνεται πλήρως από την τιμή που προκύπτει από την εξίσωση Bellman, δηλ. ο πράκτορας λαμβάνει υπόψη μόνο τις πιο πρόσφατες πληροφορίες. Συνήθως, οι τιμές του συντελεστή αυτού μειώνονται κατά τη διάρκεια της εκπαίδευσης, ώστε να επιτευχθεί μία καλή ισορροπία μεταξύ της εκμάθησης και της σταθερότητας του αλγορίθμου.

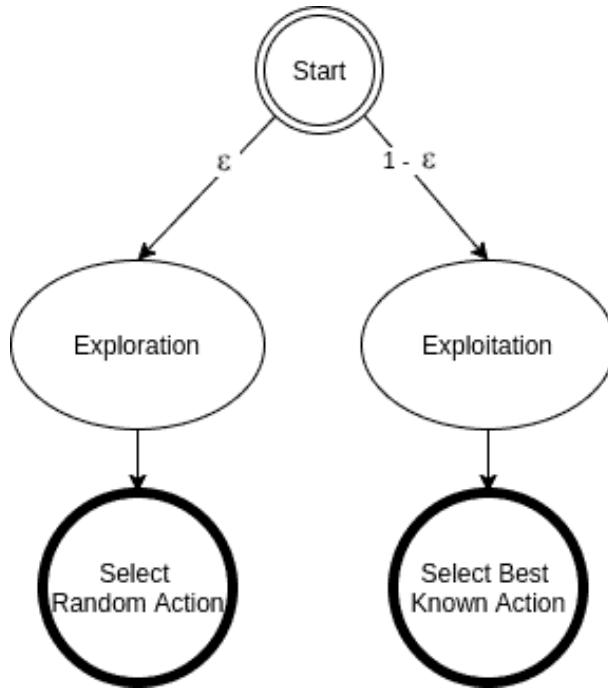
Τέλος, αφού ο πράκτορας κατασκευάσει τη συνάρτηση $Q(s, a)$, μέσα από πολλές ενημερώσεις των τιμών Q , τότε μπορεί να ενεργεί βέλτιστα μέσα από τον υπολογισμό $a = \arg \max_a Q(s, a)$, δηλ. επιλέγοντας απλώς την ενέργεια με τη μεγαλύτερη τιμή Q σε κάθε κατάσταση.

Η τεχνική ε -greedy

Ένα σημαντικό κομμάτι της παραπάνω διαδικασίας είναι το πως ο πράκτορας επιλέγει την ενέργεια που θα εκτελέσει σε κάθε κατάσταση, όταν βρίσκεται ακόμα στο στάδιο της εκπαίδευσης. Πρόκειται για το δίλημμα εξερεύνησης-αξιοποίησης, το οποίο αναλύθηκε στην παράγραφο 3.3.2. Ο αλγόριθμος Q -Learning αντιμετωπίζει το πρόβλημα αυτό, χρησιμοποιώντας την τεχνική ε -greedy.

Η τεχνική ε -greedy είναι μία διαδεδομένη μέθοδος για την εξισορρόπηση της εξερεύνησης και της αξιοποίησης. Είναι εύκολη στην υλοποίηση, αλλά ταυτόχρονα αποδίδει εξίσου καλά με άλλες, πιο πολύπλοκες μεθόδους. Η τεχνική αυτή παρουσιάζεται υπό μορφή διαγράμματος ροής (*flowchart*) στην Εικόνα 3.10.

Σημαντικό ρόλο στην τεχνική αυτή παίζει η παράμετρος ε , η οποία λαμβάνει τιμές στο διάστημα $[0, 1]$ και συμβολίζει την πιθανότητα ο πράκτορας να εξερευνήσει το περιβάλλον. Συγκεκριμένα, κατά τη διάρκεια της εκπαίδευσης, όταν ο πράκτορας βρίσκεται σε μία κατάσταση s , επιλέγει με πιθανότητα ε , τυχαία μία από τις διαθέσιμες ενέργειες (εξερεύνηση), ενώ με πιθανότητα $1 - \varepsilon$, επιλέγει την ενέργεια που έχει τη μεγαλύτερη τιμή $Q(s, a)$ στον πίνακα εκείνη τη στιγμή (αξιοποίηση).



Εικόνα 3.10. Η τεχνική ϵ -greedy (Baeldung 2023).

Συνήθως, επιλέγεται η παράμετρος ϵ μειώνεται σταδιακά κατά την εκπαίδευση κι έτσι, η τεχνική ονομάζεται και ως φθίνουσα (*decaying*) ϵ -greedy. Με τον τρόπο αυτό δίνεται μεγαλύτερη έμφαση στην εξερεύνηση στα πρώτα στάδια της εκπαίδευσης και στην αξιοποίηση στα τελευταία. Με άλλα λόγια, ο πράκτορας δρα σε μεγάλο βαθμό τυχαία στην αρχή της εκπαίδευσης, όπου η τιμή του ϵ είναι υψηλή, ενώ όταν η τιμή του ϵ φτάσει το 0, ο πράκτορας γίνεται «άπληστος» (*greedy*), δηλ. επιλέγει πάντα την βέλτιστη ενέργεια. Ο ρυθμός με τον οποίο μειώνεται η τιμή της πιθανότητας είναι συνήθως είτε γραμμικός είτε εκθετικός και εξαρτάται από τον αριθμό των καταστάσεων και των ενεργειών.

Θεωρητικά, σε άπειρο χρόνο εκπαίδευσης του αλγορίθμου *Q*-Learning, η πολιτική του πράκτορα θα αντιστοιχεί στη βέλτιστη (Melo 2001). Έτσι, δεν θα είχε τότε, νόημα η εξερεύνηση στο περιβάλλον. Ωστόσο, σε πραγματικά προβλήματα ενισχυτικής μάθησης, ο πράκτορας εκπαιδεύεται για πεπερασμένο αριθμό βημάτων, κι επομένως η τιμή ϵ δεν τείνει ποτέ στο μηδέν. Αντίθετα, τίθεται ένα κατώτατο όριο, για παράδειγμα η τιμή 0.005. Με αυτόν τον τρόπο, εξασφαλίζεται ότι ο πράκτορας δεν θα σταματήσει ποτέ, πλήρως, την εξερεύνηση, μιας και η πολιτική του θα έχει πάντα περιθώρια βελτίωσης.

Δεν υπάρχει συγκεκριμένος κανόνας για τον καθορισμό του ρυθμού μείωσης του ϵ , ούτε για το κατώτατο όριο του. Αντίθετα, ορίζονται από τον σχεδιαστή του συστήματος, μετά από πειραματισμό και δοκιμές. Είναι σημαντική η προσεκτική ρύθμιση αυτών των παραμέτρων, διότι επηρεάζουν σημαντικά τη σύγκλιση του αλγορίθμου.

Η μεθοδολογία του αλγορίθμου *Q*-Learning δίνεται παράκατω υπό μορφή ψευδοκώδικα (*pseudocode*).

Q-Learning Algorithm

```

1: Initialize  $Q(s, a)$  to 0 for all  $a \in A$  in each  $s \in S$ 
2: Initialize learning rate  $\alpha \in (0, 1]$ 
3: Initialize discount factor  $\gamma \in [0, 1]$ 
4: Initialize exploration rate  $\epsilon \in [0, 1]$ 
5: while not converged do
6:    $s \leftarrow s_0$ 
7:   while  $s$  not terminal do
8:     Observe current state  $s$ 
9:     if explore() then
10:        $a \leftarrow$  random action
11:     else
12:        $a \leftarrow \arg \max_a Q(s, a)$ 
13:     end if
14:     Take action  $a$ , observe reward  $R(s, a)$  and next state  $s'$ 
15:     Update Q-value:

$$Q(s, a) \leftarrow (1 - \alpha) \cdot Q(s, a) + \alpha \left[ R(s, a) + \gamma \cdot \max_a Q(s', a) \right]$$

16:    $s \leftarrow s'$ 
17: end while
18: end while

```

Συνολικά, η διαδικασία που ακολουθεί ο αλγόριθμος *Q*-Learning για την κατασκευή της συνάρτησης Q , θα γίνει καλύτερα κατανοητή με ένα παράδειγμα. Για αυτό, θα θεωρησούμε το παιχνίδι Frozen Lake της βιβλιοθήκης OpenAI Gymnasium², που αποτελεί ένα κλασικό περιβάλλον εφαρμογής του αλγορίθμου *Q*-Learning.

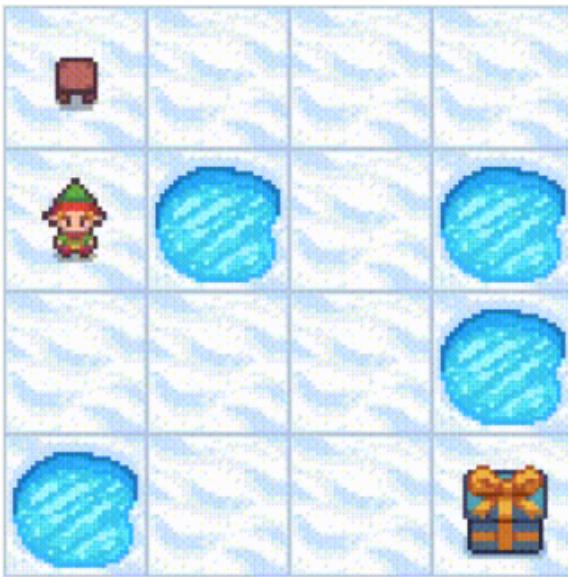
Παράδειγμα: Το παιχνίδι Frozen Lake

Στο παιχνίδι της παγωμένης λίμνης, η πίστα είναι ενα πλέγμα 4x4 και στόχος του πράκτορα είναι να φτάσει από την αρχή της πίστας (τετράγωνο πάνω αριστερά) στον στόχο του (τετράγωνο κάτω δεξιά), αποφεύγοντας τις τρύπες. Αν ο πράκτορας πέσει σε μία τρύπα, τότε το επεισόδιο τερματίζει και ο πράκτορας χάνει. Ο πράκτορας μπορεί να κινηθεί προς τα πάνω, κάτω, αριστερά και δεξιά. Ένα στιγμιότυπο του παιχνιδιού φαίνεται στην Εικόνα 3.11.

Μοντελοποιώντας το παιχνίδι σε πρόβλημα ενισχυτικής μάθησης, μπορούμε να θεωρήσουμε ως κατάσταση του περιβάλλοντος τη θέση του πράκτορα εντός του πλέγματος. Έτσι, υπάρχουν

²Η βιβλιοθήκη OpenAI Gymnasium αποτελεί μία από τις πιο δημοφιλείς βιβλιοθήκες για την εφαρμογή αλγορίθμων ενισχυτικής μάθησης σε περιβάλλοντα προσομοίωσης. Περιλαμβάνει μία μεγάλη ποικιλία από περιβάλλοντα, όπως το Frozen Lake, το CartPole, παιχνίδια Atari κ.ά. Επιπλέον, παρέχει μία εύχρηστη διεπαφή (API) για την αλληλεπίδραση με τα περιβάλλοντα, καθώς και πολλές χρήσιμες συναρτήσεις για την εκπαίδευση των αλγορίθμων.

3.3 Ενισχυτική Μάθηση



Εικόνα 3.11. Το παιχνίδι Frozen Lake της βιβλιοθήκης OpenAI Gymnasium.

16 δυνατές καταστάσεις. Επίσης, όπως αναφέραμε προηγουμένως, ο πράκτορας έχει 4 δυνατές ενέργειες. Η συνάρτηση ανταμοιβής μπορεί να εκφραστεί ως εξής:

- +1 εάν ο πράκτορας φτάσει το στόχο
- -1 εάν ο πράκτορας πέσει σε τρύπα
- 0 σε όλες τις άλλες περιπτώσεις

Στην αρχή της εκπαίδευσης, οι τιμές του πίνακα Q αρχικοποιούνται στο 0. Ο πράκτορας επιλέγει ενέργειες χρησιμοποιώντας την τεχνική ε -greedy και ενημερώνει σε κάθε βήμα την αντίστοιχη τιμή Q του ζεύγους κατάστασης-ενέργειας, χρησιμοποιώντας την εξίσωση TD Update Rule. Μετά από 50000 επεισόδια, παρατηρείται πως το ποσοστό επιτυχίας του πράκτορα προσεγγίζει το 100%. Τότε, η εκπαίδευση σταματάει και ξεκινάει η αξιολόγηση του πράκτορα, όπου τίθεται η τιμή 0 στην παράμετρο ε , ώστε να επιλέγει ο πράκτορας πάντα την βέλτιστη ενέργεια, σύμφωνα με την πολιτική του. Στην Εικόνα 3.12 παρουσιάζεται η πολιτική που ανέπτυξε ο πράκτορας (αριστερά), καθώς και η τελική μορφή του πίνακα Q (δεξιά).

Τα βέλη στο πλέγμα αριστερά αντιστοιχούν στην ενέργεια που επιλέγει στην αντίστοιχη κατάσταση ο πράκτορας. Όπως αποδεικνύεται κι από τον πίνακα δεξιά, η ενέργεια που επιλέγεται κάθε φορά είναι αυτή με τη μεγαλύτερη τιμή Q . Ακόμα, αξίζει να σημειωθεί πως οι μόνες καταστάσεις με μηδενικές τιμές Q είναι οι τερματικές καταστάσεις, αφού σε αυτές ο πράκτορας δεν εκτελεί καμία ενέργεια, καθώς το παιχνίδι τελειώνει. Τελικά, η διαδρομή που σχηματίζεται από τις επιλεγμένες ενέργειες, οδηγεί τον πράκτορα από την αφετηρία στον στόχο, κι έτσι η εκπαίδευση θεωρείται επιτυχής.



Εικόνα 3.12. Αποτελέσματα εκπαίδευσης στο παιχνίδι Frozen Lake (Szymanski 2018).

Αδυναμίες και πιθανές λύσεις

Ο αλγόριθμος *Q*-Learning αποτελεί έναν από τους θεμελιώδεις και πιο γνωστούς αλγορίθμους στον χώρο της ενισχυτικής μάθησης. Παρόλα αυτά, η χρήση του τα τελευταία χρόνια έχει μειωθεί σημαντικά, κάτι το οποίο οφείλεται στις σημαντικές αδυναμίες που εμφανίζει. Ωστόσο, πριν αναλύσουμε αυτές, ας επισημάνουμε πρώτα τα πλεονεκτήματά του αλγορίθμου, τα οποία τον κατέστησαν τόσο δημοφιλή:

- **Εύκολη υλοποίηση:** Η απλή φύση του αλγορίθμου καθιστά την υλοποίησή του αρκετά εύκολη, χωρίς να χρειάζεται η χρήση κάποιας βιβλιοθήκης μηχανικής μάθησης.
- **Μικρές απαιτήσεις σε υπολογιστικούς πόρους:** μπορεί να εκτελεστεί σε συστήματα με μικρή υπολογιστική ισχύ, σε αντίθεση με τους πιο μοντέρνους αλγορίθμους, που συνηθώς απαιτούν ειδικό πόρους όπως εξελιγμένες μονάδες επεξεργασίας γραφικών υπολογισμών (GPU).
- **Κατάλληλος για προβλήματα με μακροπρόθεσμα αποτελέσματα:** Χάρη στην ενημέρωση των τιμών *Q* με την εξίσωση TD Update Rule, ο αλγόριθμος είναι ικανός να αντιμετωπίσει προβλήματα με μακροπρόθεσμα αποτελέσματα, τα οποία είναι ιδιαίτερα απαιτητικά.

Στη συνέχεια, ας εξετάσουμε τις σημαντικότερες αδυναμίες του αλγορίθμου *Q*-Learning:

- **Ακατάλληλος για μεγάλους χώρους καταστάσεων και ενεργειών:** Σε προβλήματα με συνεχή χώρο καταστάσεων ή και ενεργειών, ο αλγόριθμος *Q*-Learning είναι πρακτικά, ανεφάρμοστος.

3.3 Ενισχυτική Μάθηση

Αυτό συμβαίνει, διότι ο αλγόριθμος απαιτεί την αποθήκευση των τιμών Q σε έναν πίνακα. Σε περιβάλλοντα όμως με μεγάλους χώρους καταστάσεων και ενεργειών, το πλήθος των διαθέσιμων ζευγών κατάστασης-ενέργειας είναι τεράστιο, με αποτέλεσμα ο πίνακας Q να γίνεται υπερβολικά μεγάλος και να απαιτεί μεγάλα ποσά μνήμης. Τα προβλήματα αυτά, που εμφανίζονται σε χώρους μεγάλων καταστάσεων είναι γνωστά και με το όνομα *κατάρα της διάστασης* (*curse of dimensionality*), ένας όρος που δόθηκε από τον Richard E. Bellman.

- **Έλλειψη γενίκευσης:** Η απλή μέθοδος Q-Learning δεν είναι ικανή να γενικεύσει τη γνώση που αποκτά. Με άλλα λόγια, ο πράκτορας δεν μπορεί να πάρει αποφάσεις για καταστάσεις που δεν έχει συναντήσει ξανά, ανεξάρτητα από την ομοιότητα της νέας αυτής κατάστασης με άλλες που έχει ήδη συναντήσει. Επομένως, σε προβλήματα με μεγάλο χώρο καταστάσεων, ακόμα και αν είχαμε τους απαραίτητους υπολογιστικούς πόρους για την αποθήκευση του πίνακα Q , ο χρόνος που θα χρειαζόταν για να εξερευνήσει ο πράκτορας κάθε ένα ζεύγος κατάστασης-ενέργειας θα ήταν απαγορευτικά μεγάλος. Τελικά, η αδυναμία γενίκευσης της γνώσης αποτελεί άλλον έναν παράγοντα για τον οποίο δεν είναι εφικτή η αποτελεσματική λειτουργία του αλγορίθμου σε πραγματικά προβλήματα.

Γίνεται πλέον κατανοητό, ότι η χρήση του αλγορίθμου Q-Learning σε πραγματικά προβλήματα είναι περιορισμένη, λόγω των παραπάνω αδυναμιών. Ωστόσο, υπάρχουν κάποιες προτάσεις για την αντιμετώπιση των προβλημάτων του αλγορίθμου. Οι κυριότερες από αυτές -και οι οποίες δοκιμάστηκαν στο περιβάλλον αυτόματης στάθμευσης- είναι οι εξής:

- **Διακριτοποίηση:** Μία τεχνική για τη μείωση του χώρου καταστάσεων και ενεργειών είναι η διακριτοποίηση των πιθανών τιμών. Για παράδειγμα, στο περιβάλλον αυτόματης στάθμευσης, αν η ταχύτητα του αυτοκινητού παίρνει συνεχείς τιμές στο διάστημα [0, 100], μπορεί να κβαντοποιηθεί στις τιμές [0, 1, 2, 3, 4], όπου το 0 αντιστοιχεί στην τιμή 0-20, το 1 στην τιμή 21-40 κ.ο.κ. Αν πραγματοποιηθεί αυτή η διαδικασία σε κάθε ένα χαρακτηριστικό του χώρου καταστάσεων, τότε το πλήθος των στοιχείων του πίνακα Q μειώνεται σημαντικά. Έτσι, μπορεί να γίνει εφικτή η εφαρμογή του αλγορίθμου Q-Learning. Βέβαια, αν η διακριτοποίηση γίνει σε υπερβολικό βάθμο, μπορεί να προκαλέσει την έλλειψη ακρίβειας στην αναπαράσταση των καταστάσεων και να οδηγήσει σε χαμηλές επιδόσεις του πράκτορα.
- **Προσέγγιση συνάρτησης:** Μία άλλη τεχνική για την αντιμετώπιση του προβλήματος της μεγάλης διάστασης του πίνακα Q είναι η χρήση συναρτήσεων προσέγγισης. Συγκεκριμένα, μπορούν να χρησιμοποιηθούν νευρωνικά δίκτυα για να προσεγγίζουν τη συνάρτηση $Q(s, a)$ και να εκτιμούν τις τιμές Q , χωρίς να χρειάζεται αυτές να αποθηκεύονται σε πίνακα. Έτσι, τα νευρωνικά δίκτυα είναι ικανά να γενικεύουν τη γνώση που αποκτούν, δηλ. να παίρνουν αποφάσεις για καταστάσεις που δεν έχουν συναντήσει ξανά αυτούσιες, κάτι που ο απλός πίνακας Q δεν μπορεί να κάνει. Η χρήση νευρωνικών δικτύων στον αλγόριθμο Q-Learning οδηγεί στη δημιουργία του αλγορίθμου Deep Q-Network (DQN). Ο αλγόριθμος αυτός αποτελεί μία από τις πιο δημοφιλείς εκδοχές του αλγορίθμου Q-Learning, καθώς είναι ικανός να

αντιμετωπίσει προβλήματα με μεγάλους, συνεχείς χώρους καταστάσεων.

Η μέθοδος της προσέγγισης συνάρτησης αποτελεί σήμερα, την πιο συνηθισμένη λύση στο πρόβλημα της διαστασιμότητας. Με τον τρόπο αυτό, εισερχόμαστε στο πεδίο της Βαθιάς Ενισχυτικής Μάθησης, το οποίο περιγράφεται στην επόμενη ενότητα.

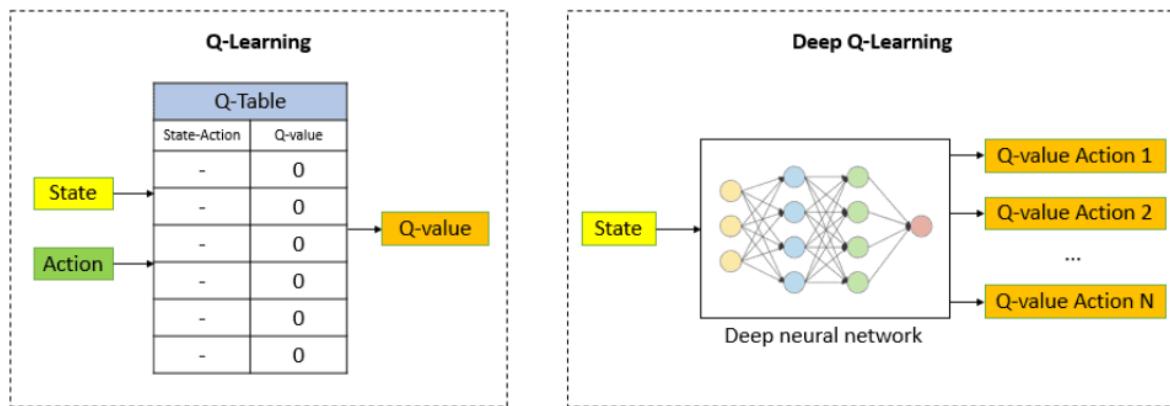
3.4 Βαθιά Ενισχυτική Μάθηση

3.4.1 Ορισμός και Χαρακτηριστικά

Τα προβλήματα της σύγχρονης εποχής χαρακτηρίζονται από υψηλή πολυπλοκότητα και μεγάλο πλήθος διαφορετικών καταστάσεων και ενεργειών. Η μεγάλη διαστασιμότητα των προβλημάτων αυτών, έχει θέσει σημαντικούς περιορισμούς στην εφαρμογή των αλγορίθμων της ενισχυτικής μάθησης. Συγκεκριμένα, σε τέτοιες περιπτώσεις, η εκπαίδευση και η ικανοποιητική εξερεύνηση του περιβάλλοντος καθίσταται χρονοβόρα, κοστοβόρα και εν τέλει, απαγορευτική από παραδοσιακούς αλγορίθμους ενισχυτικής μάθησης.

Η λύση δίνεται μέσω της μίμησης ενός βιολογικού μηχανισμού, των νευρωνικών δικτύων, και της ενσωμάτωσης τους στους αλγορίθμους ενισχυτικής μάθησης. Έτσι, δημιουργήθηκε ένα νέο επιστημονικό πεδίο, αυτό της Βαθιάς Ενισχυτικής Μάθησης. Επομένως, η Βαθιά Ενισχυτική Μάθηση (*Deep Reinforcement Learning*) αποτελεί απλά την ενοποίηση της Βαθιάς Μάθησης (*Deep Learning* - χρήση βαθιών νευρωνικών δικτύων) με την Ενισχυτική Μάθηση (*Reinforcement Learning*).

Μία σύγκριση της κλασικής ενισχυτικής μάθησης με τη βαθιά ενισχυτική μάθηση παρουσιάζεται στην *Εικόνα 3.13*, για την περίπτωση του αλγορίθμου *Q-Learning*.



Εικόνα 3.13. Κλασική εναντίον Βαθιάς Ενισχυτικής Μάθησης (Quang 2024).

Παρατηρώντας την *Εικόνα 3.13* γίνεται σαφής η λογική της βαθιάς ενισχυτικής μάθησης καθώς και τα πλεονεκτήματα που προκύπτουν από την χρήση νευρωνικών δικτύων. Συγκεκριμένα, στον κλασικό

3.4 Βαθιά Ενισχυτική Μάθηση

αλγόριθμο Q-Learning, όλες οι τιμές Q των ζευγών κατάστασης ενέργειας αποθηκεύονται σε έναν πίνακα. Ο πράκτορας ανατρέχει στον πίνακα αυτό, για να επιλέξει σε κάθε κατάσταση, την ενέργεια με τη μεγαλύτερη αξία. Αντίθετα, ο αλγόριθμος βαθιάς ενισχυτικής μάθησης DQN, αντικαθιστά τον προηγούμενο πίνακα με ένα βαθύ νευρωνικό δίκτυο. Το δίκτυο αυτό, εκτιμά σε κάθε κατάσταση, τις τιμές Q για όλες τις δυνατές ενέργειες, ώστε να επιλέξει ο πράκτορας τη βέλτιστη ενέργεια. Επομένως, αποφεύγεται η αποθήκευση των τιμών Q , ενώ επιτυγχάνεται η γενίκευση σε άγνωστες καταστάσεις.

Τα τεχνητά νευρωνικά δίκτυα αποτελούν το αντικείμενο της επόμενης παραγράφου, όπου αρχικά συγκρίνονται με τα βιολογικά νευρωνικά δίκτυα, ενώ στη συνέχεια παρουσιάζονται οι βασικές αρχές λειτουργίας τους.

3.4.2 Τεχνητά Νευρωνικά Δίκτυα

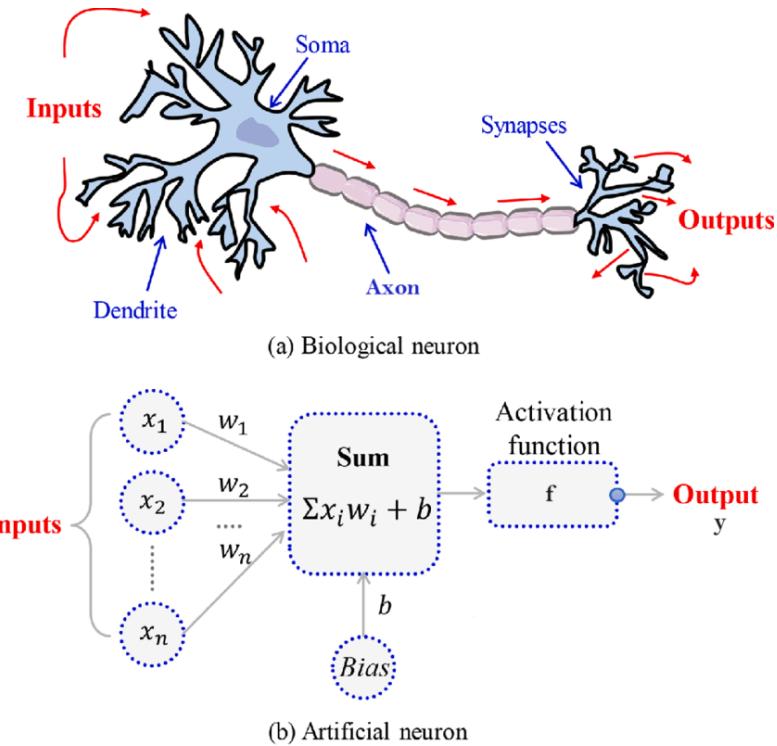
Τα Τεχνητά Νευρωνικά Δίκτυα (*Artificial Neural Networks - ANNs*) είναι υπολογιστικά μοντέλα εμπνευσμένα από τη δομή και τη λειτουργία των βιολογικών νευρωνικών δικτύων που βρίσκονται στον ανθρώπινο εγκέφαλο. Αποτελούνται από συνδεδεμένα επίπεδα κόμβων (ή *νευρώνων*) που συνεργάζονται για να επεξεργαστούν και να μάθουν από τα δεδομένα.

Σύγκριση με Βιολογικά Νευρωνικά Δίκτυα

Όπως προαναφέρθηκε, τα τεχνητά νευρωνικά δίκτυα είναι σχεδιασμένα ώστε να μιμούνται τη λειτουργία των βιολογικών νευρωνικών δικτύων. Προκεμένου να κατανοήσουμε πώς λειτουργούν τα νευρωνικά δίκτυα, ας εξετάσουμε πρώτα πώς λειτουργούν τα βιολογικά νευρωνικά δίκτυα. Ο ανθρώπινος εγκέφαλος αποτελείται από δισεκατομμύρια κύτταρα που ονομάζονται νευρώνες, οι οποίοι επικοινωνούν μεταξύ τους μέσω ηλεκτρικών και χημικών σημάτων. Η Εικόνα 3.14 δείχνει μια σύγκριση μεταξύ ενός βιολογικού νευρώνα και ενός τεχνητού νευρώνα, επισημαίνοντας τα δομικά τους στοιχεία.

Από το πάνω μέρος της εικόνας 3.14, παρατηρούμε πως ένας βιολογικός νευρώνας αποτελείται από τέσσερις κύριες μονάδες: τους δενδρίτες, το σώμα, τους άξονες και τις συνάψεις. Οι δενδρίτες λαμβάνουν τα σήματα εισόδου και τα μεταφέρουν στο σώμα, όπου επεξεργάζονται. Έπειτα, οι άξονες μεταφέρουν το επεξεργασμένο σήμα σε άλλους νευρώνες μέσω συνάψεων. Οι συνάψεις λειτουργούν ως σύνδεσμοι ανάμεσα στους νευρώνες.

Στον τεχνητό νευρώνα, παρατηρούμε πως οι είσοδοι x_1, x_2, \dots, x_n παίζουν το ρόλο των δενδριτών, ενώ τα βάρη w_1, w_2, \dots, w_n αντιστοιχούν στις συνάψεις. Ο υπολογισμός του σταθμισμένου αθροίσματος $z = \sum_{i=1}^n w_i x_i$ αντιστοιχεί στην επεξεργασία του σήματος στο σώμα του νευρώνα. Τέλος, η έξοδος y αντιστοιχεί στο σήμα που μεταφέρεται μέσω των αξόνων. Οι έννοιες αυτές αναλύονται λεπτομερώς στη συνέχεια.



Εικόνα 3.14. Σύγκριση βιολογικού και τεχνητού νευρώνα (Karpathy 2016).

Αρχές Λειτουργίας

Ας ξεκινήσουμε εξετάζοντας τη λειτουργία ενός μεμονωμένου νευρώνα, σε ένα τεχνητό νευρωνικό δίκτυο. Κάθε νευρώνας λαμβάνει εισόδους από άλλους νευρώνες, οι οποίες συμβολίζονται ως x_i . Κάθε είσοδος πολλαπλασιάζεται με έναν συντελεστή, που ονομάζεται βάρος (*weight*) και συμβολίζεται ως w_i . Αξίζει να σημειωθεί εδώ, πως τα βάρη παίζουν κρίσιμο ρόλο στη λειτουργία του νευρωνικού δικτύου. Συγκεκριμένα, μεγάλα βάρη υποδηλώνουν ότι οι συγκεκριμένες μεταβλητές είναι πιο σημαντικές για την τελική απόφαση του δικτύου. Επομένως, τα βάρη αυτά πρέπει να προσαρμοστούν κατά την εκπαίδευση του δικτύου, ώστε να αντικατοπτρίζουν την πραγματική σημασία των μεταβλητών.

Στη συνέχεια, υπολογίζεται το σταθμισμένο άθροισμα όλων των εισόδων $z = \sum_{i=1}^n w_i x_i$. Στο άθροισμα αυτό, προστίθεται κι ένας άλλος παράγοντας, ο οποίος ονομάζεται πόλωση (*bias*) και συμβολίζεται ως b . Ο παράγοντας αυτός αποτελεί μία σταθερά και χρησιμοποιείται για την οριζόντια μετατόπιση της εξόδου του νευρώνα. Με αυτόν τον τρόπο, το bias βελτιώνει την προσαρμογή του μοντέλου στα δεδομένα και την ακρίβεια του.

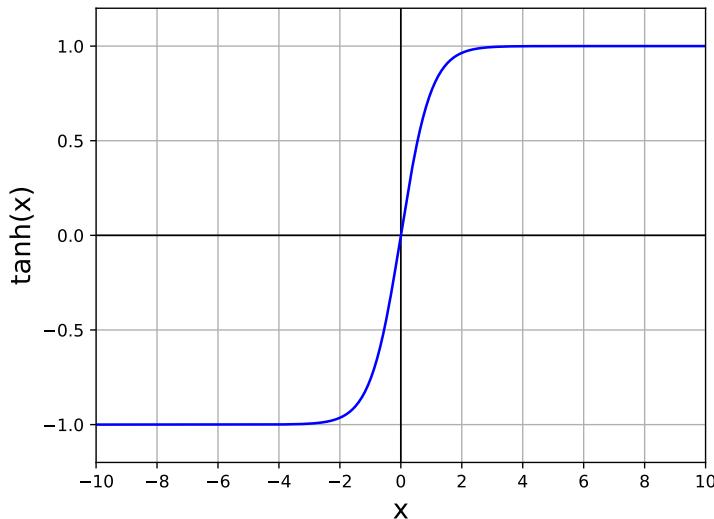
Τέλος, το προηγούμενο αποτέλεσμα περνάει από μια συνάρτηση ενεργοποίησης, για να προκύπτει η τελική έξοδος του νευρώνα. Η συνάρτηση ενεργοποίησης (*activation function*) είναι μία μαθηματική

3.4 Βαθιά Ενισχυτική Μάθηση

συνάρτηση, η οποία συμβολίζεται ως f και εφαρμόζεται στην έξοδο του νευρώνα, πριν όμως αυτή περάσει στο επόμενο επίπεδο του δικτύου. Η συνάρτηση ενεργοποίησης αποφασίζει αν η έξοδος είναι αρκετά σημαντική, για να περάσει στο επόμενο επίπεδο. Συγκεκριμένα, αν η έξοδος υπερβαίνει ένα καθορισμένο κατώφλι, θα περάσει και λέμε ότι ο νευρώνας «ενεργοποιείται». Οι συναρτήσεις ενεργοποίησης αποτελούν ένα πολύ σημαντικό στοιχείο των νευρωνικών δικτύων, για δύο λόγους. Πρώτον, εισάγουν μη γραμμικότητα (*non linearity*) στο δίκτυο, η οποία είναι εξαιρετικά σημαντική, καθώς χάρη σε αυτήν, καθίσταται δυνατή η επίλυση πολύπλοκων προβλημάτων από τα νευρωνικά δίκτυα. Δεύτερον, ελέγχουν το εύρος των εξόδων των νευρώνων, το οποίο μπορεί να είναι κρίσιμο για τη σταθεροποίηση και την επιτάχυνση της διαδικασίας μάθησης. Υπάρχουν διάφορες συναρτήσεις ενεργοποίησης, με διαφορετικά χαρακτηριστικά η κάθε μία. Η επιλογή της συνάρτησης ενεργοποίησης εξαρτάται από παράγοντες όπως οι απαιτήσεις μάθησης του δικτύου και οι ιδιαιτερότητες του εκάστοτε προβλήματος. Στην πράξη, η επιλογή γίνεται συνήθως μετά από δοκιμές και πειραματισμούς, καθώς διαφορετικές συναρτήσεις ενεργοποίησης μπορεί να επιδρούν διαφορετικά στην απόδοση του δικτύου. Δύο από τις πιο διάσημες συναρτήσεις ενεργοποίησης -και οι οποίες χρησιμοποιήθηκαν στο πρόβλημα αυτόματης στάθμευσης- παρουσιάζονται παρακάτω:

- **Συνάρτηση Tanh:** Η συνάρτηση υπερβολικής εφαπτομένης (*Hyperbolic Tangent - Tanh*) ορίζεται από την εξίσωση 3.4 και σχεδιάζεται στην Εικόνα 3.15.

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \text{όπου } x \text{ είναι η έξοδος του νευρώνα \quad (3.4)}$$



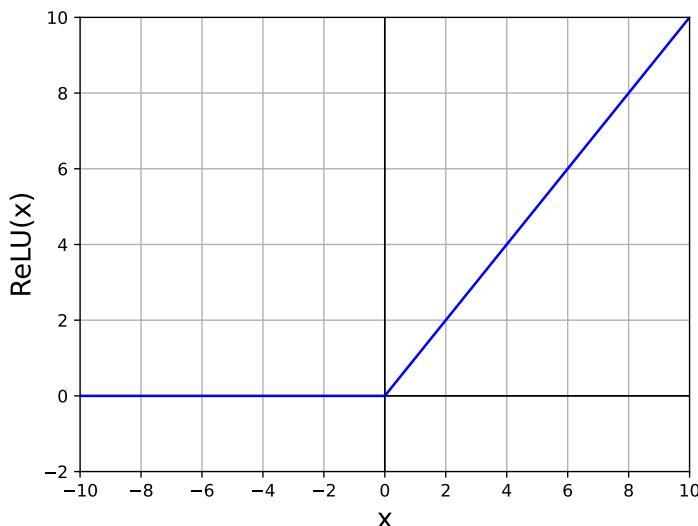
Εικόνα 3.15. Γραφική παράσταση της συνάρτησης Tanh.

Από την Εικόνα 3.15 παρατηρούμε πως το εύρος εξόδου της συνάρτησης Tanh είναι $[-1, 1]$. Έτσι, η

έξοδος της Tanh είναι συμμετρική, με κέντρο το 0. Αυτό είναι χρήσιμο, καθώς μπορεί να επιταχύνει τη σύγκλιση του αλγορίθμου κατά την εκπαίδευση. Ακόμα, χάρη σε αυτήν την ίδιοτητα της, όταν τα δεδομένα εισόδου είναι επίσης κανονικοποιημένα ώστε να έχουν μέση τιμή 0, η εκπαίδευση μπορεί να γίνει πιο αποδοτική. Το βασικό μειονέκτημα της συνάρτησης Tanh είναι το πρόβλημα της εξαφάνισης της κλίσης (*vanishing gradients problem*). Το πρόβλημα αυτό, εμφανίζεται όταν η είσοδος της συνάρτησης γίνεται πολύ μικρή ή πολύ μεγάλη. Τότε, η κλίση της συνάρτησης προσεγγίζει το 0. Αυτό οδηγεί σε πολύ μικρές αλλαγές στα βάρη κατά την εκπαίδευση, κι επομένως η μάθηση του δικτύου γίνεται πολύ αργή ή ακόμη και σταματά.

- Συνάρτηση ReLU: Η συνάρτηση Ανορθωμένης Γραμμικής Μονάδας (*Rectified Linear Unit - ReLU*) ορίζεται από την εξίσωση 3.5 και σχεδιάζεται στην Εικόνα 3.16.

$$f(x) = \max(0, x), \text{όπου } x \text{ είναι η έξοδος του νευρώνα \quad (3.5)}$$

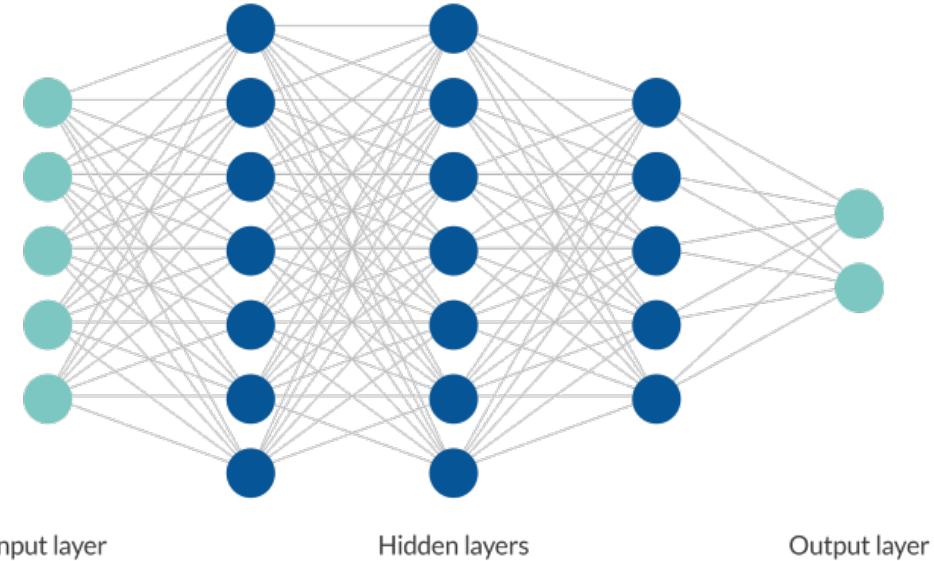


Εικόνα 3.16. Γραφική παράσταση της συνάρτησης ReLU.

Από την Εικόνα 3.16 παρατηρούμε πως το εύρος εξόδου της συνάρτησης ReLU είναι $[0, \infty)$. Άρα, δεν υπάρχει ανώτατο όριο για την έξοδο της συνάρτησης κι επομένως, η κανονικοποίηση των δεδομένων εισόδου είναι ξανά μία καλή πρακτική. Η συνάρτηση ReLU αποτελεί την πιο ευρέως χρησιμοποιούμενη συνάρτηση ενεργοποίησης. Είναι δημοφιλής για την μικρή απαιτούμενη υπολογιστική ισχύ, για την επιτάχυνση της σύγκλισης και την αποφυγή του προβλήματος της εξαφάνισης της κλίσης. Ωστόσο, έχει κάποια μειονεκτήματα, με κυριότερο την τάση νεκρών νευρώνων (*dying ReLU problem*), όπου κάποιοι νευρώνες σταματούν να ανταποκρίνονται σε οποιαδήποτε είσοδο και παραμένουν ανενεργοί.

3.4 Βαθιά Ενισχυτική Μάθηση

Πλέον, έχοντας μελετήσει τη λειτουργία ενός μεμονωμένου νευρώνα, ας εξετάσουμε πως συνεργάζονται πολλοί νευρώνες για να δημιουργήσουν ένα νευρωνικό δίκτυο. Στην *Εικόνα 3.17* παρουσιάζεται ένα παράδειγμα ενός τεχνητού νευρωνικού δικτύου.



Εικόνα 3.17. Παράδειγμα τεχνητού νευρωνικού δικτύου (Comsol 2023).

Παρατηρούμε πως τα νευρωνικά δίκτυα αποτελούνται από πολλά επίπεδα (ή στρώματα) νευρώνων, τα οποία συνδέονται μεταξύ τους. Το πρώτο επίπεδο, ονομάζεται επίπεδο εισόδου (*input layer*) και λαμβάνει τα δεδομένα εισόδου του δικτύου. Τα δεδομένα αυτά στη συνέχεια μεταφέρονται μέσα από μία σειρά επιπέδων που ονομάζονται ενδιάμεσα επίπεδα (*hidden layers*) και είναι υπεύθυνα για την επεξεργασία τους και την εξαγωγή χαρακτηριστικών από αυτά. Τέλος, τα δεδομένα φτάνουν στο τελευταίο επίπεδο, το οποίο ονομάζεται επίπεδο εξόδου (*output layer*) και παράγει την έξοδο του δικτύου.

Ανάλογα με την πολυπλοκότητα του προβλήματος, ένα νευρωνικό δίκτυο μπορεί να αποτελείται από ένα ή περισσότερα ενδιάμεσα επίπεδα. Όταν το δίκτυο αποτελείται από δύο ή περισσότερα ενδιάμεσα επίπεδα, τότε ονομάζεται βαθύ νευρωνικό δίκτυο (*deep neural network*). Αντίστοιχα, το κάθε επίπεδο μπορεί να αποτελείται από λίγους έως και εκατομμύρια νευρώνες.

Ακόμα, ένα νευρωνικό δίκτυο μπορεί να διαφέρει στον τρόπο με τον οποίο συνδέονται οι νευρώνες του. Η πιο διαδεδομένη τοπολογία είναι αυτή των πλήρως συνδεδεμένων επιπέδων (*fully connected layers*). Σε αυτήν, ο κάθε νευρώνας ενός επιπέδου συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου.

Εκπαίδευση

Τα τεχνητά νευρωνικά δίκτυα μαθαίνουν μέσω της προσαρμογής των βαρών και των πολώσεων τους κατά τη διάρκεια της εκπαίδευσης. Ας εξετάσουμε τη διαδικασία αυτή βήμα προς βήμα, αναλύοντας παράλληλα ορισμένες σημαντικές έννοιες και παραμέτρους που εμπλέκονται σε αυτήν.

Στο 1^o βήμα της εκπαίδευσης γίνεται η αρχικοποίηση των βαρών, δηλ. η ανάθεση τυχαίων τιμών σε αυτά.

Στο 2^o βήμα της εκπαίδευσης λαμβάνει χώρα η εμπρόσθια διάδοση, κατά την οποία τα δεδομένα εισόδου περνάνε μέσα από το δίκτυο, για να παραχθεί η έξοδος του.

Στο 3^o βήμα της εκπαίδευσης, υπολογίζεται το σφάλμα της εξόδου, δηλ. η διαφορά της εξόδου του δίκτυου από την επιθυμητή τιμή της εξόδου³. Ο υπολογισμός γίνεται από την **συνάρτηση σφάλματος** (*loss function*) και στόχος είναι, μέσα από την εκπαίδευση, να ελαχιστοποιηθεί αυτό το σφάλμα. Μερικές από τις πιο διάσημες συναρτήσεις σφάλματος είναι η συνάρτηση Μέσης Τετραγωνικής Απόκλισης (*Mean Squared Error - MSE*), η συνάρτηση Μέσου Απόλυτου Σφάλματος (*Mean Absolute Error - MAE*) και η συνάρτηση σφάλματος Εντροπίας (*Cross-Entropy*) και η επιλογή της εξαρτάται από τον τύπο του προβλήματος.

Στο 4^o βήμα της εκπαίδευσης λαμβάνει χώρα η **οπισθοδρόμικη διάδοση** του σφάλματος (*backward propagation* ή *backpropagation*). Κατά τη συγκεκριμένη διαδικασία, υπολογίζεται η κλίση (δηλ. η μερική παράγωγος) της συνάρτησης σφάλματος ως προς κάθε βάρος του δίκτυου. Ο υπολογισμός αυτός γίνεται προς τα πίσω, δηλ. από το επίπεδο εξόδου προς το επίπεδο εισόδου και αντικατοπτρίζει την επίδραση του κάθε βάρους στο σφάλμα της εξόδου. Αξίζει να σημειωθεί πως, από τεχνικής άποψης, η οπισθοδρόμηση αποτελεί απλώς τη μέθοδο για τον αποτελεσματικό υπολογισμό της κλίσης της συνάρτησης σφάλματος και δεν έχει σχέση με το πως χρησιμοποιείται αυτή στη συνέχεια. Ωστόσο, συχνά, αυτός ο όρος χρησιμοποιείται ελαστικά, για να περιγράψει τη συνολική διαδικασία μάθησης.

Στο 5^o βήμα της εκπαίδευσης, γίνεται η ενημέρωση των βαρών του δίκτυου. Η ενημέρωση των βαρών γίνεται σύμφωνα με τις υπολογισμένες μερικές παραγώγους, προς την κατεύθυνση που μειώνει το σφάλμα και πραγματοποιείται από έναν **αλγόριθμο βελτιστοποίησης** (*optimization algorithm*). Οι πιο συνηθισμένοι αλγόριθμοι βελτιστοποίησης είναι:

- Ο αλγόριθμος **Κατάβασης Πλαγιάς** (*Gradient Descent*): είναι ο πιο απλός αλγόριθμος βελτιστοποίησης, αλλά παρουσιάζει προβλήματα όπως την αργή σύγκλιση και την πιθανότητα να παγιδευτεί σε τοπικά ελάχιστα. Αποτέλεσε τη βάση για την ανάπτυξη πιο προηγμένων αλγορίθμων, ενώ υπάρχουν και πολλές παραλλαγές του, όπως ο αλγόριθμος **Στοχαστικής**

³σε προβλήματα επιβλεπόμενης μάθησης, η επιθυμητή τιμή της εξόδου είναι γνωστή για τα δεδομένα εκπαίδευσης. Σε προβλήματα ενισχυτικής μάθησης, ο υπολογισμός του σφάλματος γίνεται με βάση την ανταμοιβή που λαμβάνει το δίκτυο από το περιβάλλον. Περισσότερες λεπτομέρειες στην παράγραφο «Σύνδεση με την Ενισχυτική Μάθηση».

3.4 Βαθιά Ενισχυτική Μάθηση

Κατάβασης Πλαγιάς (*Stochastic Gradient Descent - SGD*) και ο αλγόριθμος Κατάβασης Πλαγιάς Τεμαχισμένου Ρυθμού (*Mini-batch Gradient Descent*).

- Ο αλγόριθμος **Ορμής** (*Momentum*): αποτελεί επέκταση του SGD, προσθέτοντας έναν όρο «օρμής», δηλ. ένα τμήμα της προηγούμενης ενημέρωσης στην τρέχουσα ενημέρωση. Βοηθάει στην επιτάχυνση της σύγκλισης και τη μείωση των ταλαντώσεων.
- Ο αλγόριθμος **Διάδοσης Μέσης Τετραγωνικής Απόκλισης** (*Root Mean Square Propagation - RMSprop*): προσαρμόζει τον ρυθμό μάθησης για κάθε παράμετρο ξεχωριστά και κανονικοποιεί το βήμα της ενημέρωσης. Αυτό βοηθάει στην αντιμέτωπιση του προβλήματος της εξαφάνισης της κλίσης (*vanishing gradients*).
- Ο αλγόριθμος **Αδάμ** (*Adam*): συνδυάζει τα πλεονεκτήματα των αλγορίθμων Momentum και RMSprop. Προσαρμόζει το ρυθμό μάθησης για κάθε παράμετρο ξεχωριστά, βασιζόμενος σε προηγούμενες κλίσεις. Είναι πολύ αποδοτικός, καθώς επιτυγχάνει ταχεία σύγκλιση και απαιτεί μικρή ποσότητα μνήμης. Στην πράξη, οι επιδόσεις του Adam συχνά ξεπερνούν τους υπόλοιπους αλγορίθμους βελτιστοποίησης και για αυτό, αποτελεί μία από τις πιο δημοφιλείς επιλογές σήμερα.

Μία σημαντική παράμετρος των αλγορίθμων βελτιστοποίησης είναι ο **ρυθμός μάθησης** (*learning rate*), ο οποίος καθορίζει το πόσο γρήγορα τα βάρη του δικτύου προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης. Η προσαρμογή του ρυθμού μάθησης είναι κρίσιμη για να επιτευχθεί ισορροπία μεταξύ της ταχύτητας σύγκλισης και της αποφυγής της υπερπήδησης (*overshooting*) της βέλτιστης λύσης.

Επίσης, αξίζει να σημειωθεί, πως στα πλαίσια αυτής της εργασίας, έγινε η επιλογή του αλγορίθμου βελτιστοποίησης Adam, χάρη στην ικανότητα του να επιτυγχάνει καλά αποτελέσματα, γρήγορα.

Τέλος, τα βήματα 2-5 επαναλαμβάνονται για κάθε τεμάχιο (*batch*) των δεδομένων εκπαίδευσης. Ένα πλήρες πέρασμα όλων των δεδομένων εκπαίδευσης ονομάζεται **εποχή** (*epoch*). Μπορεί να χρειαστούν πολλές εποχές εκπαίδευσης, ώστε να επιτευχθεί η επιθυμητή απόδοση του δικτύου.

Υπερ-παράμετροι

Οι παράμετροι του νευρωνικού δίκτυου που ορίζονται πριν την εκπαίδευση από τον σχεδιαστή του, ονομάζονται **υπερ-παράμετροι** (*hyperparameters*). Οι παράμετροι αυτοί, έχουν πολύ μεγάλη επίδραση στην τελική αποτελεσματικότητα του δικτύου κι επομένως επιβάλλεται η προσεκτική επιλογή και ρύθμιση τους. Δεν υπάρχουν συγκεκριμένοι κανόνες που να ορίζουν την επιλογή των υπερ-παραμέτρων, κι έτσι αυτή πρέπει να γίνει μέσω δοκιμών και κατάλληλων ρυθμίσεων. Αξίζει να σημειωθεί, πως οι αλλαγές των υπερ-παραμέτρων δεν μπορούν να λάβουν χώρα κατά τη διάρκεια της εκπαίδευσης. Έτσι, αν κριθεί απαραίτητη η αλλαγή κάποιας παραμέτρου του δικτύου, η εκπαίδευση πρέπει να επαναληφθεί από την αρχή. Μερικές από τις βασικότερες υπερ-παραμέτρους ενός νευρωνικού δικτύου είναι:

- **Η αρχιτεκτονική του δικτύου.** Ο όρος «αρχιτεκτονική» ή «δομή» του δικτύου περιλαμβάνει τον αριθμό των επιπέδων, τον αριθμό των νευρώνων σε κάθε επίπεδο καθώς και τη συνδεσμολογία μεταξύ τους. Συνήθως, η αύξηση της πολυπλοκότητας του δικτύου, δηλ. του πλήθους των επιπέδων και των νευρώνων, βελτιώνει την ικανότητα του να μαθαίνει σύνθετα μοτίβα δεδομένων και αυξάνει έτσι την απόδοση του. Ωστόσο, αυξάνοντας την πολυπλοκότητα του δικτύου, αυξάνεται και ο κίνδυνος της υπερ-προσαρμογής του στα δεδομένα εκπαίδευσης (*overfitting*), με αποτέλεσμα να μην γενικεύει καλά σε νέα δεδομένα (ΕΙΤCA 2024).
- **Η συνάρτηση ενεργοποίησης:** Η επιλογή της συνάρτησης ενεργοποίησης επηρεάζει την ικανότητα του δικτύου να μάθει και την ταχύτητα της μάθησης.
- **Ο ρυθμός μάθησης:** Η επιλογή ενός υψηλού ρυθμού μάθησης μπορεί να οδηγήσει σε αστάθεια τη διαδικασία εκπαίδευσης, ενώ από την άλλη, ένας χαμηλός ρυθμός μάθησης μπορεί να καθυστερήσει τη σύγκλιση του δικτύου.
- **Ο αλγόριθμος βελτιστοποίησης:** Η επιλογή του αλγορίθμου βελτιστοποίησης επηρεάζει την ταχύτητα σύγκλισης και την σταθερότητα της εκπαίδευσης.

Προφανώς, η λίστα αυτή δεν είναι εξαντλητική. Υπάρχουν πολλές ακόμα υπερπαράμετροι στο εσωτερικό των δικτύων αυτών και κάποιος μπορεί να αναλύσει τη λειτουργία τους σε πολύ μεγαλύτερο βάθος. Ωστόσο, θεωρώ πως οι παραπάνω παράμετροι αποτελούν αφενός μία καλή αφετηρία για έναν αρχάριο στο αντικείμενο και αφετέρου, στα πλαίσια χρήσης των νευρωνικών δικτύων για την επίλυση προβλημάτων ενισχυτικής μάθησης, θεωρώ πως αποτελούν αυτές με τις οποίες αξίζει να πειραματιστεί πρώτα κάποιος.

Κατηγορίες

Υπάρχουν διαφορετικοί τύποι νευρωνικών δικτύων, οι οποίοι διακρίνονται από την αρχιτεκτονική τους και από τον τύπο προβλημάτων που είναι σχεδιασμένοι να επιλύσουν. Οι πιο κοινοί τύποι νευρωνικών δικτύων σήμερα είναι οι εξής:

- **Τα Νευρωνικά Δίκτυα Πρόσθιας Διάδοσης (*Feedforward Neural Networks*):** Αποτελούν την πιο βασική μορφή νευρωνικών δικτύων. Όπως γίνεται φανερό και από το όνομα τους, χαρακτηριστικό αυτών των δικτύων είναι πως τα δεδομένα εισόδου διαδίδονται μόνο προς την κατεύθυνση της έξοδου, χωρίς να υπάρχουν κύκλοι ή πλάγιες συνδέσεις. Μάλιστα, όταν τα δίκτυα αυτά αποτελούνται από πολλά επίπεδα, πλήρως συνδεδεμένα μεταξύ τους -δηλ. στην πλειοψηφία των περιπτώσεων-, τότε ονομάζονται και **Πολυεπίπεδα Νευρωνικά Δίκτυα (*Multilayer Perceptrons - MLPs*)**. Παρά την απλή δομή τους, τα ενδιάμεσα επίπεδα των δικτύων αυτών μπορούν να είναι πολύπλοκα κι έτσι, τα μοντέλα αυτά χρησιμοποιούνται σε διάφορες εργασίες, όπως προβλήματα ταξινόμησης και παλινδρόμησης.
- **Τα Συνελικτικά Νευρωνικά Δίκτυα (*Convolutional Neural Networks - CNNs*):** Τα δίκτυα αυτά χρησιμοποιούν την πράξη της συνέλιξης για την επεξεργασία δεδομένων. Είναι ιδιαίτερα

3.4 Βαθιά Ενισχυτική Μάθηση

αποτελεσματικά στην αναγνώριση προτύπων ή εικόνων και την ανίχνευση αντικειμένων. Έτσι, καθίστανται εξαιρετικά χρήσιμα σε εφαρμογές υπολογιστικής όρασης. Τα CNNs αποτελούνται από τρία βασικά είδη επιπέδων: τα επίπεδα συνέλιξης, υπεύθυνα για την εξαγωγή χαρακτηριστικών από την είσοδο, τα επίπεδα υποδειγματοληψίας, υπεύθυνα για τη μείωση της διάστασης των προηγούμενων χαρακτηριστικών και τα πλήρως συνδεδεμένα επίπεδα, υπεύθυνα για την τελική ταξινόμηση των δεδομένων.

- Τα **Ανατροφοδοτούμενα Νευρωνικά Δίκτυα** (*Recurrent Neural Networks - RNNs*): Τα δίκτυα αυτά διακρίνονται από τους κύκλους ανάδρασης τους, επιτρέποντας έτσι την αποθήκευση πληροφορίας εντός του δικτύου. Με τον τρόπο αυτό, τα RNNs διατηρούν μία μνήμη των προηγούμενων εισόδων και μπορούν να χειριστούν επιτυχώς ακολουθιακά δεδομένα, δηλ. περιπτώσεις όπου η εισερχόμενη πληροφορία είναι διαδοχικής φύσεως. Για παράδειγμα, χρησιμοποιούνται ευρέως σε εφαρμογές όπως η αναγνώριση ομιλίας, η πρόβλεψη χρονοσειρών και η επεξεργασία φυσικής γλώσσας.
- Τα **Παραγωγικά Δίκτυα Αντιπαλότητας** (*Generative Adversarial Networks - GANs*): Η βασική διαφοροποίηση των δικτύων αυτών, είναι πως ουσιαστικά, αποτελούνται από δύο ξεχωριστά νευρωνικά δίκτυα. Το πρώτο δίκτυο ονομάζεται «παραγωγός» (*generator*) και είναι υπεύθυνο για τη δημιουργία νέων εικόνων ή κειμένου με βάση ένα σύνολο δεδομένων εκπαίδευσης. Το δεύτερο δίκτυο ονομάζεται «κρίτης» και στόχος του είναι να κρίνει το έργο του γεννήτορα, αποφασίζοντας εάν φαίνεται πραγματικό ή ψεύτικο. Η εκπαίδευση ολοκληρώνεται όταν ο κριτής δεν μπορεί να διακρίνει μεταξύ των δεδομένων εκπαίδευσης και των έργων του γεννήτορα. Έτσι, τα GANs χρησιμοποιούνται για την παραγωγή νέου περιεχομένου, όπως κείμενο, εικόνες και βίντεο.

Τέλος, αξίζει να σημειώθει πως οι έννοιες που αναλύθηκαν στις προηγούμενες παραγράφους περιγράφουν τη λειτουργία των πολυεπίπεδων νευρωνικών δικτύων (*MLPs*). Οι υπόλοιποι τύποι νευρωνικών δικτύων βασίζονται στις ίδιες αρχές και τρόπο λειτουργίας, αλλά, όπως είδαμε, έχουν και επιπλέον, ειδικές ιδιότητες και μηχανισμούς. Επιλέχθηκε να αναλυθεί πιο συγκεκριμένα η κατηγορία των MLPs, επειδή αποτελούν όχι μόνο τη βάση για την κατανόηση των γενικών αρχών των νευρωνικών δικτύων, αλλά και τον τύπο που χρησιμοποιήθηκε στα πλαίσια αυτής της εργασίας.

Σύνδεση με την Ενισχυτική Μάθηση

Έχοντας μελετήσει τις θεμελιώδεις αρχές και τον τρόπο λειτουργίας των τεχνητών νευρωνικών δικτύων, ας εξετάσουμε την χρήση τους σε προβλήματα ενισχυτικής μάθησης. Η κεντρική ιδέα είναι πως ο πράκτορας εκφράζεται ως ένα βαθύ νευρωνικό δίκτυο, το οποίο δέχεται ως είσοδο την κατάσταση του περιβάλλοντος και επιστρέφει την επιλεγμένη ενέργεια. Κατά τη διάρκεια της εκπαίδευσης, αναπροσαρμόζονται τα βάρη του νευρωνικού δικτύου, με στόχο την επίτευξη της μέγιστης συνολικής ανταμοιβής.

Ωστόσο, υπάρχουν μερικές διαφοροποιήσεις όσον αφορά την έξοδο του δικτύου και τον τρόπο εκπαίδευσής του, με βάση την κατηγορία του αλγορίθμου ενισχυτικής μάθησης που χρησιμοποιείται. Επομένως, ας εξετάσουμε πιο αναλυτικά τις διαφορές αυτές για τις 3 κατηγορίες model free αλγορίθμων που αναλύσαμε στην παράγραφο 3.3.3.

Αλγόριθμοι Εκτίμησης Αξίας

Οι πιο χρησιμοποιούμενοι αλγόριθμοι αυτής της κατηγορίας είναι οι αλγόριθμοι Εκτίμησης Αξίας Ζευγών Κατάστασης-Ενέργειας (αλγόριθμοι Q-Learning), όπως πχ ο αλγόριθμος DQN. Οι αλγόριθμοι αυτοί εφαρμόζονται σε περιβάλλοντα με διακριτό χώρο ενέργειών και χρησιμοποιούν ένα δίκτυο αξίας για την εκτίμηση των τιμών Q των ζευγών κατάστασης-ενέργειας.

- Έξοδος Δικτύου: η έξοδος του νευρωνικού δικτύου αποτελείται από τις τιμές Q για όλες τις διαθέσιμες ενέργειες στην τρέχουσα κατάσταση του περιβάλλοντος. Η επιλογή ενέργειας γίνεται με βάση κάποια τεχνική όπως η ϵ -greedy, για την αντιμέτωπιση του διλήμματος εξερεύνησης-αξιοποίησης, όπως έχει αναφερθεί στην παράγραφο 3.3.4.
- Εκπαίδευση Δικτύου: η εκπαίδευση του δίκτυου γίνεται με βάση την ανταμοιβή που λαμβάνει ο πράκτορας από το περιβάλλον. Συγκεκριμένα, η συνάρτηση σφάλματος που χρησιμοποιείται ονομάζεται Σφάλμα Χρονικών Διαφορών (*Temporal Difference Error - TD Error*), είναι η διαφορά μεταξύ της εκτιμώμενης τιμής Q του δικτύου και της πραγματικής τιμής Q . Η πραγματική τιμή Q υπολογίζεται από την εξίσωση Bellman (βλ. 3.2), χρησιμοποιώντας την τρέχουσα ανταμοιβή του πράκτορα από το περιβάλλον.

Αλγόριθμοι Βελτιστοποίησης Πολιτικής

Σε αυτούς τους αλγορίθμους χρησιμοποιείται ένα δίκτυο πολιτικής, δηλ. το νευρωνικό δίκτυο αντικατοπτρίζει απευθείας την πολιτική του πράκτορα.

- Έξοδος Δικτύου:
 - Διακριτός Χώρος Ενέργειών: η έξοδος του δικτύου είναι μία κατανομή πιθανοτήτων στις διαθέσιμες ενέργειες πχ $P(\text{action}_1) = 0.5$, $P(\text{action}_2) = 0.4$, $P(\text{action}_3) = 0.1$. Η επιλογή ενέργειας εξαρτάται από τον χαρακτηρισμό της πολιτικής ως ντετερμινιστική ή στοχαστική. Εάν η πολιτική είναι ντετερμινιστική, τότε θα επιλεγεί η ενέργεια με τη μεγαλύτερη πιθανότητα. Αντίθετα, εάν η πολιτική είναι στοχαστική, τότε η επιλογή της ενέργειας θα γίνει μέσω δειγματοληψίας από την κατανομή πιθανοτήτων.
 - Συνεχής χώρος Ενέργειών: η έξοδος του δικτύου, για κάθε ενέργεια, αποτελείται από 2 παραμέτρους μίας Γκαουσιανής κατανομής: τη μέση τιμή της ενέργειας (μ) και την τυπική απόκλιση της (σ). Για παράδειγμα, αν οι ενέργειες του πράκτορα ήταν η ταχύτητα και η γωνία ενός αυτοκινήτου, τότε η έξοδος του δικτύου θα ήταν της μορφής: $\mu(\text{velocity})$, $\sigma(\text{velocity})$, $\mu(\text{angle})$, $\sigma(\text{angle})$. Η επιλογή της συγκεκριμένης τιμής κάθε ενέργειας, εξαρτάται ξανά από την πολιτική του δικτύου. Αν η πολιτική είναι ντετερμινιστική,

τότε θα επιλεγεί απλώς η μέση τιμή της ενέργειας. Αντίθετα, αν η πολιτική είναι στοχαστική, τότε η επιλογή της τιμής θα γίνει μέσω δειγματοληψίας από την κατανομή που περιγράφεται από τη μέση τιμή και την τυπική απόκλιση.

- Εκπαίδευση Δικτύου: η εκπαίδευση του δίκτυου γίνεται με βάση την ανταμοιβή που λαμβάνει ο πράκτορας από το περιβάλλον. Συγκεκριμένα, σε κάθε βήμα t γίνεται μία εκτίμηση για την αθροιστική ανταμοιβή G_t που θα λάβει ο πράκτορας:

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots \quad (3.6)$$

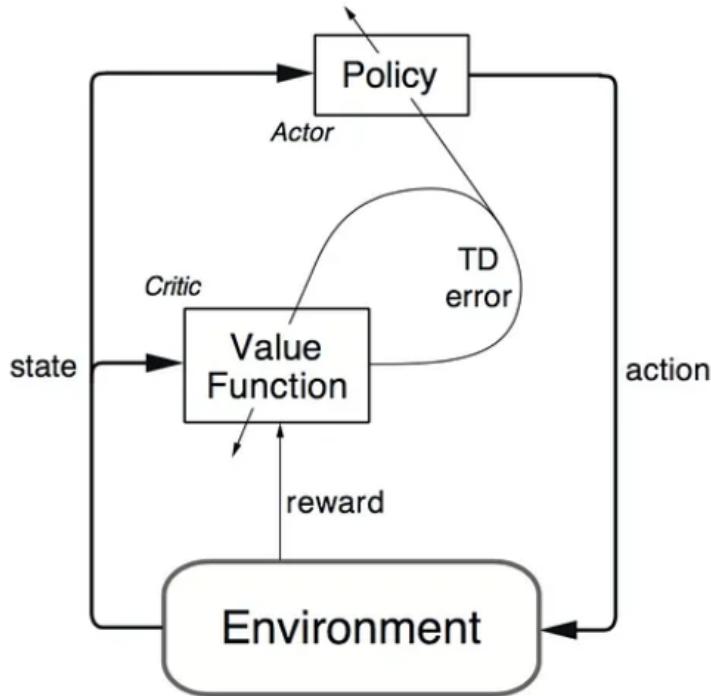
όπου R_t είναι η τρέχουσα ανταμοιβή και γ ο παράγοντας έκπτωσης. Στόχος είναι η μεγιστοποίηση της συνάρτησης G_t κι έτσι, υπολογίζεται η κλίση της ως προς τις παραμέτρους θ της πολιτικής. Η κλίση αυτή χρησιμοποιείται για την ενημέρωση των βαρών του δικτύου.

Αλγόριθμοι Δράστη-Κριτή

Στους αλγορίθμους αυτούς χρησιμοποιούνται δύο νευρωνικά δίκτυα, ο δράστης, υπεύθυνος για την επιλογή της ενέργειας και ο κριτής, υπεύθυνος για την εκτίμηση της αξίας της τρέχουσας κατάστασης ή της επιλεγμένης ενέργειας στην τρέχουσα κατάσταση.

- Έξοδος Δικτύων:
 - Δράστης: η έξοδος του δράστη αποτελεί την πολιτική του πράκτορα. Για την έξοδο του δικτύου και την τελική επιλογή ενέργειας, ισχύουν τα όσα περιγράφησαν νωρίτερα, στους αλγορίθμους βελτιστοποίησης πολιτικής.
 - Κριτής: η έξοδος του κριτή αποτελεί την εκτίμηση μίας αξίας. Σε ορισμένους αλγορίθμους, αυτή η αξία αναφέρεται στην τρέχουσα κατάσταση $V(s)$, ενώ σε άλλους αλγορίθμους αναφέρεται στο ζεύγος τρέχουσας κατάστασης-επιλεγμένης ενέργειας $Q(s, a)$.
- Εκπαίδευση Δικτύων: η εκπαίδευση γίνεται και σε αυτήν την περίπτωση, με βάση την ανταμοιβή του περιβάλλοντος. Η διαδικασία αποτυπώνεται παραστατικά στην *Εικόνα 3.18*. Συγκεκριμένα, βλέπουμε πως ο κριτής δέχεται ως είσοδο την τρέχουσα κατάσταση και ανταμοιβή του περιβάλλοντος. Έτσι, υπολογίζει το σφάλμα χρονικών διαφορών (*TD Error*), όπως περιγράφηκε νωρίτερα, στους αλγορίθμους εκτίμησης αξίας. Το σφάλμα χρονικών διαφορών χρησιμοποιείται για την ενημέρωση των βαρών του κριτή και του δράστη, όπως φαίνεται από τις καμπύλες που διαπερνούν τα δίκτυα αυτά στην *Εικόνα 3.18*. Με τον τρόπο αυτό, ο κριτής εκπαιδεύει το δίκτυο του, ενώ παρέχει και ανάδραση στον δράστη, η οποία χρησιμοποιείται για την εκπαίδευση του.

Έχοντας πλέον, αναλύσει τα βασικά χαρακτηριστικά της βαθιάς ενισχυτικής μάθησης, καθώς και τον τρόπο λειτουργίας των νευρωνικών δικτύων σε αυτό το πλαίσιο, ας εξετάσουμε τους αλγορίθμους



Εικόνα 3.18. Αλληλεπίδραση δικτύων Δράστη-Κριτή (Sutton και Barto 2018).

βαθιάς ενισχυτικής μάθησης που χρησιμοποιήθηκαν στην παρούσα εργασία. Συγκεκριμένα, πραγματοποιήθηκαν εκπαιδεύσεις πρακτόρων με τους εξής αλγόριθμους:

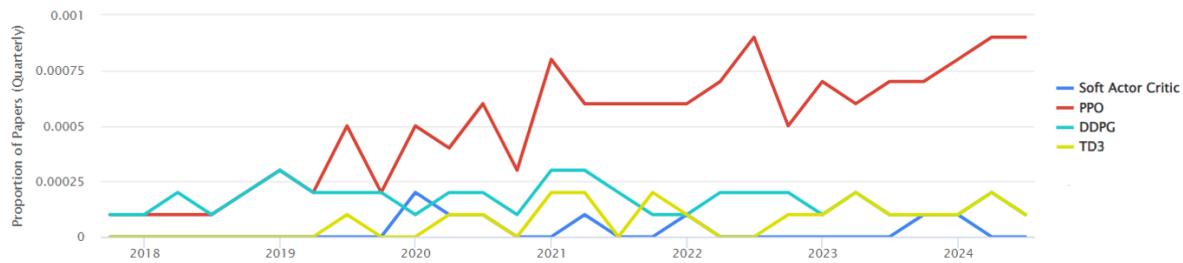
- PPO
- SAC
- TD3
- DDPG

3.4.3 Ο αλγόριθμος PPO

Ο αλγόριθμος PPO (*Proximal Policy Optimization*) είναι ένας αποτελεσματικός αλγόριθμος ενισχυτικής μάθησης, που ανήκει στην κατηγορία των αλγορίθμων βελτιστοποίησης πολιτικής. Αναπτύχθηκε το 2017 από την εταιρία OpenAI (Schulman κ.ά. 2017) και ξεχωρίζει για την απλότητά του, τη σταθερότητα του και τις ισχυρές επιδόσεις που πετυχαίνει. Σήμερα, αποτελεί έναν από τους πιο δημοφιλείς αλγορίθμους ενισχυτικής μάθησης. Αυτό αποδεικνύεται και από την Εικόνα 3.19, όπου παρουσιάζεται η συχνότητα των επιστημονικών δημοσιεύσεων των αλγορίθμων ενισχυτικής μάθησης που χρησιμοποιήθηκαν στην παρούσα εργασία, σε σχέση με τον χρόνο. Παρατηρούμε πως ο αλγόριθμος PPO έχει πολύ μεγαλύτερη συχνότητα χρήσης, σε σχέση με τους υπόλοιπους αλγορίθμους.

3.4 Βαθιά Ενισχυτική Μάθηση

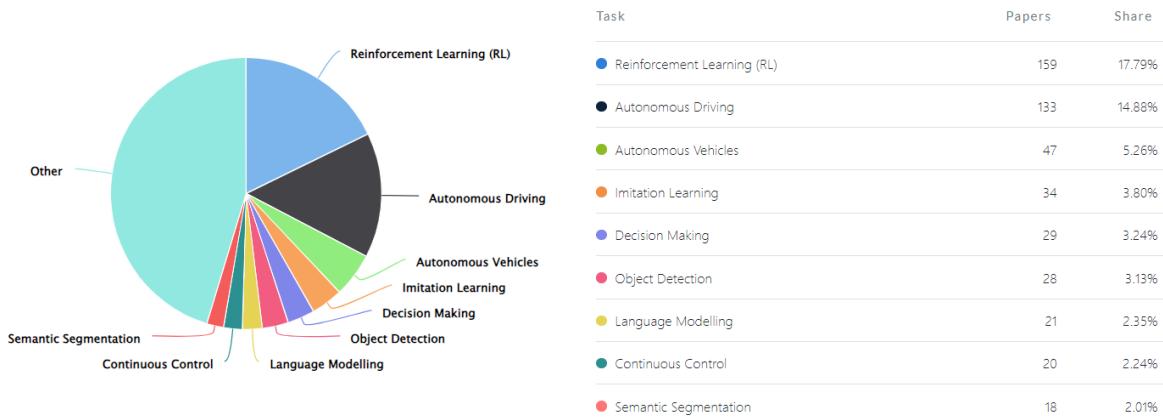
Usage Over Time



Εικόνα 3.19. Συχνότητα χρήσης αλγορίθμων ενισχυτικής μάθησης σε επιστημονικές δημοσιεύσεις (Papers With Code 2024).

Επίσης, στην *Εικόνα 3.20* παρουσιάζονται οι τύποι των προβλημάτων στα οποία χρησιμοποιείται ο αλγόριθμος PPO. Παρατηρούμε πως ο πιο συνήθης τύπος είναι τα προβλήματα ενισχυτικής μάθησης, ακολουθούμενα από τα προβλήματα αυτόνομης οδήγησης και αυτόνομων οχημάτων. Άρα, καταλαβαίνουμε πως ο αλγόριθμος PPO θα είναι κατάλληλος, για την επίλυση του προβλήματος αυτόματης στάθμευσης που αντιμετωπίζουμε στην παρούσα εργασία.

Tasks



Εικόνα 3.20. Τύποι προβλημάτων αλγορίθμου PPO [Papers With Code (2024)]⁴.

Ο PPO βελτιστοποιεί την πολιτική του πράκτορα, μέσω της μεγιστοποίησης μιας αντικειμενικής συνάρτησης. Η αντικειμενική συνάρτηση αυτή είναι «περικομμένη» (*clipped*), δηλ. έχει ένα κατώφλι που περιορίζει την αλλαγή της πολιτικής σε κάθε βήμα. Έτσι, αποφεύγονται μεγάλες, απότομες

⁴ Αξίζει να σημειωθεί πως τα δύο προηγούμενα γραφήματα προέρχονται από τη σελίδα [paperswithcode](#), η οποία παρέχει χρήσιμες πληροφορίες σχετικά με αλγορίθμους μηχανικής μάθησης, όπως επιστημονικές δημοσιεύσεις, κώδικα, αποτελέσματα κ.α.

αλλαγές της πολιτικής και διασφαλίζεται η σταθερότητα της εκπαίδευσης. Η αντικειμενική συνάρτηση του αλγορίθμου PPO παρουσιάζεται στην εξίσωση 3.7:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (3.7)$$

Οι όροι της αντικειμενικής συνάρτησης αναλύονται παρακάτω:

- $r_t(\theta)$ αποτελεί το λόγο της νέας πολιτικής προς την παλιά και ισούται με $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$
- \hat{A}_t είναι η συνάρτηση πλεονεκτήματος, η οποία εκτιμά το πόσο καλύτερη είναι μια ενέργεια σε μία συγκεκριμένη κατάσταση σε σχέση με τη μέση ενέργεια
- $\text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)$ είναι η συνάρτηση περικοπής, η οποία διασφαλίζει ότι η πολιτική δεν θα υποστεί δραστικές αλλαγές, περιορίζοντας το $r_t(\theta)$ στο εύρος $[1 - \epsilon, 1 + \epsilon]$.
- ϵ είναι το κατώφλι που περιορίζει την αλλαγή της πολιτικής. Αποτελεί υπερ-παράμετρο του αλγορίθμου και παίρνει μικρές θετικές τιμές (συνήθως 0.2).
- Η συνάρτηση πίν χρησιμοποιείται για την επιλογή του ελάχιστου μεταξύ της μη-περικομμένης και περικομμένης αντικειμενικής συνάρτησης. Έτσι, η τελική αντικειμενική συνάρτηση είναι ένα κάτω φράγμα (δηλ. μια απαισιόδοξη εκτίμηση) της μη-περικομμένης αντικειμενικής συνάρτησης.

Επίσης, συχνά προστίθεται κι ένας όρος εντροπίας στην αντικειμενική συνάρτηση, ο οποίος ενθαρρύνει τον πράκτορα να εξερευνήσει. Έτσι, διατηρείται μία ποικιλία στις ενέργειες που επιλέγονται από την πολιτική, στα πρώτα στάδια της εκπαίδευσης.

Η μεθοδολογία του αλγορίθμου PPO δίνεται παράκατω υπό μορφή ψευδοκώδικα.

PPO Algorithm

- 1: **Input:** initial policy parameters θ_0 , clipping threshold ϵ
- 2: **for** $k = 0, 1, 2, \dots$ **do**
- 3: Collect set of partial trajectories D_k using policy $\pi_k = \pi(\theta_k)$
- 4: Estimate advantages $\hat{A}_t^{\pi_k}$ using any advantage estimation algorithm
- 5: Compute policy update:

$$\theta_{k+1} = \arg \max_{\theta} L_{\theta_k}^{\text{CLIP}}(\theta)$$
- 6: Perform K steps of minibatch SGD (via Adam), where:

$$L_{\theta_k}^{\text{CLIP}}(\theta) = E_{\tau \sim \pi_k} \left[\sum_{t=0}^T \min \left(r_t(\theta) \hat{A}_t^{\pi_k}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t^{\pi_k} \right) \right]$$

- 7: **end for**
-

Ένα σημαντικό χαρακτηριστικό της εκπαίδευσης του αλγορίθμου PPO, είναι πως δεν εκτελεί μόνο μια ενημέρωση για κάθε δεδομένο που συλλέγεται. Αντίθετα, πραγματοποιεί πολλαπλές εποχές

3.4 Βαθιά Ενισχυτική Μάθηση

ενημερώσεων, χρησιμοποιώντας τα ίδια συλλεγμένα δεδομένα. Αυτό καθιστά τον αλγόριθμο πιο αποδοτικό στη χρήση δειγμάτων, σε σύγκριση με άλλες μεθόδους βελτιστοποίησης πολιτικής. Ακόμα, η εκπαίδευση γίνεται on-policy, ενώ χρησιμοποιείται μια στοχαστική πολιτική.

Όσον αφορά την αρχιτεκτονική του νευρωνικού δικτύου, στην αρχική δημοσίευση των (Schulman κ.ά. 2017), χρησιμοποιήθηκε ένα πλήρως συνδεδεμένο MLP με δύο κρυφά επίπεδα των 64 μονάδων και συνάρτηση ενεργοποίησης Tanh.

Με αυτούς τους τρόπους, ο PPO πετυχαίνει την ανθεκτικότητα και την αξιοπιστία παλαιοτέρων αλγορίθμων βελτιστοποίησης πολιτικής, όπως ο TRPO, αλλά είναι πολύ απλούστερος, ευκολότερος στην υλοποίηση και πετυχαίνει συνολικά, υψηλότερες επιδόσεις. Επιπλέον, είναι αποδοτικός στη χρήση δειγμάτων κι έτσι, μειώνεται το πλήθος των απαιτούμενων αλληλεπιδράσεων με το περιβάλλον. Τέλος, αξιοσημειώτη είναι η ευελιξία του αλγορίθμου, καθώς μπορεί να εφαρμοστεί τόσο σε διακριτούς, όσο και σε συνεχείς χώρους ενεργειών.

3.4.4 Ο αλγόριθμος DDPG

Ο αλγόριθμος DDPG (*Deep Deterministic Policy Gradient*) είναι ένας αλγόριθμος βαθιάς ενισχυτικής μάθησης, που ανήκει στην κατηγορία των αλγορίθμων δράστη-κριτή. Αναπτύχθηκε το 2015 από ερευνητές της εταιρίας DeepMind της Google (Lillicrap κ.ά. 2015). Ο DDPG συνδυάζει στοιχεία των αλγορίθμων βελτιστοποίησης πολιτικής, καθώς και του αλγορίθμου εκτίμησης αξίας DQN. Είναι σχεδιασμένος για περιβάλλοντα με συνεχείς χώρους καταστάσεων και ενεργειών και επομένως, μπορεί να θεωρηθεί ως η εφαρμογή του αλγορίθμου DQN σε συνεχείς χώρους ενεργειών.

Ο DDPG, χρησιμοποιεί ένα δίκτυο δράστη για την επιλογή της ενέργειας. Τα βάρη του δικτύου αυτού συμβολίζονται ως θ^{μ} . Ακόμα, όπως φανερώνει και το όνομα του αλγορίθμου, ο DDPG χρησιμοποιεί ντετερμινιστική πολιτική. Αυτό σημαίνει πως από το συνεχές εύρος τιμών της κάθε ενέργειας, ο δράστης δεν επιλέγει μία μέση τιμή $\mu(s_t)$ και μία τυπική απόκλιση $\sigma(s_t)$, αλλά επιλέγει απευθείας μία τιμή $\mu(s_t)$. Ωστόσο, ένα σημαντικό χαρακτηριστικό του αλγορίθμου DDPG, είναι πως προσθέτει θόρυβο στην ενέργεια που επιλέγει ο δράστης, για να ενθαρρύνει την εξερεύνηση του περιβάλλοντος. Ο θόρυβος αυτός συνήθως ακολουθεί κατανομή Gauss ή Ornstein-Uhlenbeck. Επομένως, η ενέργεια που επιλέγει ο δράστης δίνεται από την εξίσωση 3.8:

$$a_t = \mu(s_t | \theta^{\mu}) + N_t \quad (3.8)$$

Επιπλέον, χρησιμοποιείται ένα δίκτυο κριτή για την εκτίμηση της αξίας των ζευγών κατάστασης-ενέργειας, δηλ. των τιμών Q . Τα βάρη του δικτύου αυτού συμβολίζονται ως θ^Q .

Ένα ακόμα χαρακτηριστικό του αλγορίθμου είναι η εκπαίδευση off policy. Συγκεκριμένα, αποθηκεύονται προηγούμενες εμπειρίες της μορφής (s_t, a_t, R_t, s_{t+1}) σε μία προσωρινή μνήμη που ονομάζεται *replay buffer*. Κατά την εκπαίδευση, μερικές φορές χρησιμοποιούνται τυχαία

δείγματα από τον replay buffer αντί για τα δεδομένα που συλλέγονται εκείνη τη χρονική στιγμή. Με αυτόν τον τρόπο, περιορίζονται οι συσχετίσεις μεταξύ διαδοχικών δεδομένων και επιτυγχάνεται μεγαλύτερη σταθερότητα στην εκπαίδευση.

Μία άλλη, σημαντική ιδιότητα του αλγορίθμου DDPG είναι η κανονικοποίηση τεμαχίων (*batch normalization*). Αυτή η τεχνική χρησιμοποιείται για την κανονικοποίηση των εισόδων στα δίκτυα δράστη και κριτή, έτσι ώστε να έχουν μονάδικη μέση τιμή και διακύμανση. Αυτό βοηθά στην αύξηση της σταθερότητάς της εκπαίδευσης.

Τέλος, αξιοσημείωτη αποτελεί η χρήση δύο δικτύων στόχου (*target networks*), ένα για τον δράστη και ένα για τον κριτή. Τα δίκτυα αυτά έχουν βάρη $\theta^{\mu'}$ και $\theta^{Q'}$ αντίστοιχα, τα οποία ενημερώνονται αργά, με βάση τα βάρη των κύριων δικτύων. Μάλιστα, χρησιμοποιείται ένας παράγοντας τ που ελέγχει τον ρυθμό ενημέρωσης των βαρών των δικτύων στόχων και παίρνει μικρές τιμές (συνήθως $\tau = 0.001$). Με τον τρόπο αυτό, οι ενημερώσεις γίνονται πιο ομαλές και αποφεύγονται μεγάλες ταλαντώσεις κατά την εκπαίδευση.

Τα δίκτυα στόχου χρησιμοποιούνται στον υπολογισμό της τιμής Q στόχου (y_t), ο οποίος γίνεται μέσα από την εξίσωση 3.9:

$$y_t = R_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1} | \theta^{\mu'}) | \theta^{Q'}) \quad (3.9)$$

Επομένως, το δίκτυο του κριτή ενημερώνεται μέσω της ελαχιστοποίησης του τετραγώνου της διαιφοράς, μεταξύ της πρόβλεψης Q του κριτή και της τιμής στόχου y_t . Η αντίστοιχη συνάρτηση σφάλματος παρουσιάζεται στην εξίσωση 3.10:

$$L(\theta^Q) = \frac{1}{N} \sum_t (y_t - Q(s_t, a_t | \theta^Q))^2 \quad (3.10)$$

Αντίστοιχα, το δίκτυο του δράστη ενημερώνεται μέσω της μέγιστοποίησης της αναμενόμενης ανταμοιβής, όπως παρουσιάζεται στην εξίσωση 3.11:

$$\nabla_{\theta^{\mu}} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a | \theta^Q) \Big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^{\mu}} \mu(s | \theta^{\mu}) \Big|_{s=s_i} \quad (3.11)$$

Με βάση αυτά, η μεθοδολογία του αλγορίθμου DDPG δίνεται και στην επόμενη σελίδα, υπό μορφή ψευδοκώδικα.

Συνολικά, η προσέγγιση του αλγορίθμου DDPG είναι απλή, εύκολη στην υλοποίηση και επεκτάσιμη σε δύσκολα προβλήματα, μεγάλης διαστασιμότητας. Ακόμα, οι τεχνικές που χρησιμοποιεί, όπως τα δίκτυα στόχων και ο replay buffer, βελτιώνουν σε μεγάλο βαθμό την ανθεκτικότητα της μάθησης.

Ωστόσο, οι ίδιες τεχνικές προκαλούν και μερικές αδυναμίες του αλγορίθμου, όπως η αργή σύγκλιση και η ανάγκη για μεγάλο αριθμό επαναλήψεων. Επιπλέον, πρόκληση αποτελεί η εξερεύνηση του περιβάλλοντος, καθώς η ντετερμινιστική πολιτική του αλγορίθμου μπορεί να οδηγήσει σε τοπικά

3.4 Βαθιά Ενισχυτική Μάθηση

ελάχιστα. Ο προστιθέμενος θόρυβος βοηθάει σε αυτό το πρόβλημα, αλλά συχνά δεν αρκεί για να επιτευχθεί η απαιτούμενη εξερεύνηση.

DDPG Algorithm

```

1: Initialize critic network  $Q(s, a|\theta^Q)$  and actor  $\mu(s|\theta^\mu)$  with weights  $\theta^Q$  and  $\theta^\mu$ 
2: Initialize target networks  $Q'$  and  $\mu'$  with weights  $\theta^{Q'} \leftarrow \theta^Q$ ,  $\theta^{\mu'} \leftarrow \theta^\mu$ 
3: Initialize replay buffer  $R$ 
4: for episode = 1 to  $M$  do
5:   Initialize a random process  $N$  for action exploration
6:   Receive initial observation state  $s_1$ 
7:   for  $t = 1$  to  $T$  do
8:     Select action  $a_t = \mu(s_t|\theta^\mu) + N_t$  according to the current policy and exploration noise
9:     Execute action  $a_t$  and observe reward  $r_t$  and new state  $s_{t+1}$ 
10:    Store transition  $(s_t, a_t, r_t, s_{t+1})$  in  $R$ 
11:    Sample a random minibatch of  $N$  transitions  $(s_i, a_i, r_i, s_{i+1})$  from  $R$ 
12:    Calculate target Q-value:  $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$ 
13:    Update critic by minimizing the loss:  $L = \frac{1}{N} \sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$ 
14:    Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q) \Big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu) \Big|_{s=s_i}$$

15:    Update the target networks:

$$\theta^{Q'} \leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'}$$


$$\theta^{\mu'} \leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'}$$

16:  end for
17: end for

```

3.4.5 Ο αλγόριθμος TD3

Ο αλγόριθμος TD3 (*Twin Delayed Deep Deterministic policy gradient*) αποτελεί μία βελτιωμένη έκδοση του αλγορίθμου DDPG, η οποία αναπτύχθηκε το 2018 από ερευνητές των πανεπιστημίου του Montreal και του Amsterdam (Fujimoto, Hoof, και Meger 2018). Είναι σχεδιασμένος για περιβάλλοντα με συνεχείς χώρους καταστάσεων και ενεργειών, ενώ στόχος του είναι να αντιμετωπίσει μερικά από τα μειονεκτήματα του DDPG, όπως την υπερεκτίμηση της συνάρτησης αξίας και την αστάθεια κατά την εκπαίδευση. Ο TD3 πετυχαίνει αυτόν τον στόχο, μέσω της εισαγωγής τριών βασικών τροποποιήσεων: της εξομάλυνσης της πολιτικής στόχου (*target policy smoothing*), της χρήσης του περικομμένου Διπλού Q-learning (*clipped Double Q-learning*) και της καθυστέρησης της ενημέρωσης

της πολιτικής (*delayed policy updates*).

Όπως και ο προκάτοχός του, ο TD3 ανήκει στην κατηγορία των αλγορίθμων δράστη-κριτή. Μάλιστα, χρησιμοποιεί ένα δίκτυο δράστη και δύο δίκτυα κριτή -εξού και ο χαρακτηρισμός *Twin* (δίδυμος) στο όνομα του αλγορίθμου-. Οι παράμετροι του δικτύου δράστη συμβολίζονται ως ϕ και τα βάρη των δικτύων κριτή ως θ_1 και θ_2 .

Η πρώτη τροποποίηση του αλγορίθμου είναι η εξομάλυνση της πολιτικής στόχου (*target policy smoothing*). Η τεχνική αυτή εισάγεται στον υπολογισμό της ενέργειας στόχου \tilde{a} , η οποία θα χρησιμοποιηθεί στη συνέχεια στον υπολογισμό της τιμής Q στόχου. Συγκεκριμένα, στον υπολογισμό της \tilde{a} , προστίθεται ένας μικρός βαθμός θορύβου (ϵ), στην ενέργεια που επέλεξε ο δράστης. Ο θόρυβος περιορίζεται σε ένα εύρος $[-c, c]$, όπου c είναι μία σταθερά, ώστε η ενέργεια στόχου να μην αποκλίνει υπερβολικά από την ενέργεια που επέλεξε ο δράστης. Η διαδικασία αυτή φαίνεται στην εξίσωση 3.12:

$$\tilde{a} \sim \pi_{\phi'}(s') + \epsilon, \epsilon \sim \text{clip}(N(0, \tilde{\sigma}), -c, c) \quad (3.12)$$

Με αυτόν τον τρόπο, μειώνεται η διασπορά στις εκτιμήσεις των τιμών Q και αποτρέπεται η πολιτική από την εκμετάλλευση πιθανών αιχμών της συνάρτησης Q , που συνήθως οφείλονται σε σφάλματα υπερεκτίμησης από τα δίκτυα κριτή.

Η δεύτερη τροποποίηση του αλγορίθμου αφορά το πρόβλημα της υπερεκτίμησης των τιμών Q (*overestimation bias*). Το πρόβλημα αυτό, εμφανίζεται συχνά στους αλγορίθμους εκτίμησης αξίας και περισσότερο, όταν χρησιμοποιούνται συναρτήσεις προσέγγισης, όπως τα νευρωνικά δίκτυα. Το αποτέλεσμα του, είναι η υιοθέτηση υποβέλτιστων πολιτικών από τον πράκτορα. Τέτοια σφάλματα εμφανίζονται και στην περίπτωση των αλγορίθμων δράστη-κριτή, αφού στους συγκεκριμένους αλγόριθμους, η πολιτική ενημερώνεται με βάση την εκτίμηση των τιμών Q . Ο αλγόριθμος TD3 επιδιώκει να ελαχιστοποιήσει το πρόβλημα της υπερεκτίμησης, χρησιμοποιώντας μία περικομμένη (*clipped*) εκδοχή του αλγορίθμου *Double Q-learning*. Σύμφωνα με την εκδοχή αυτή, η τιμή Q στόχου (y) υπολογίζεται από την εξίσωση 3.13:

$$y = R + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \tilde{a}) \quad (3.13)$$

Επομένως, η τιμή Q στόχου υπολογίζεται ως το ελάχιστο από τις τιμές Q που επιστρέφουν τα δύο δίκτυα κριτή. Αυτός ο κανόνας μπορεί να προκαλέσει το αντίθετο φαινόμενο, την υπεκτίμηση των τιμών Q . Ωστόσο, αυτό είναι σαφώς προτιμότερο από την υπερεκτίμηση, καθώς οι τιμές των υποεκτιμημένων ενεργειών δεν θα μεταδοθούν κατά τη μάθηση, αφού οι ενέργειες με μικρές τιμές Q αποφεύγονται από την πολιτική.

Με βάση την τιμή Q στόχου (y), τα δίκτυα κριτή ενημερώνουν τις παραμέτρους τους θ_1 και θ_2 , μέσω της ελαχιστοποίησης του τετραγώνου της διαφοράς μεταξύ της πρόβλεψης του κριτή και της τιμής

3.4 Βαθιά Ενισχυτική Μάθηση

Q στόχου. Η αντίστοιχη συνάρτηση σφάλματος παρουσιάζεται στην εξίσωση 3.14:

$$\theta_i = \arg \min_{\theta_i} \frac{1}{N} \sum (y - Q_{\theta_i}(s, a))^2 \quad (3.14)$$

Η τελευταία τροποποίηση του TD3 είναι η καθυστέρηση της ενημέρωσης της πολιτικής. Συγκεκριμένα, το δίκτυο του δράστη (και κατ' επέκταση το δίκτυο στόχου του δράστη) δεν ενημερώνεται σε κάθε βήμα της εκπαίδευσης, αλλά ανά d βήματα (προτείνεται $d = 2$). Αυτή η καθυστέρηση βοηθά στη σταθεροποίηση της διαδικασίας μάθησης, επιτρέποντας στα δίκτυα κριτή να συγκλίνουν καλύτερα πριν γίνει η ενημέρωση της πολιτικής. Έτσι, οι λιγότερο συχνές ενημερώσεις της πολιτικής θα χρησιμοποιούν μία εκτίμηση της τιμής Q με μικρότερη διακύμανση και συνεπώς, θα αποτελούν ενημερώσεις υψηλότερης ποιότητας. Οι ενημερώσεις αυτές της πολιτικής δίνονται στην εξίσωση 3.15:

$$\nabla_{\phi} J(\phi) = \frac{1}{N} \sum \nabla_a Q_{\theta_1}(s, a) \Big|_{a=\pi_{\phi}(s)} \nabla_{\phi} \pi_{\phi}(s) \quad (3.15)$$

Με βάση τη διαδικασία που περιγράφηκε προηγουμένως, η μεθοδολογία του αλγορίθμου TD3 δίνεται παράκατω υπό μορφή ψευδοκώδικα.

TD3 Algorithm

- 1: Initialize critic networks Q_{θ_1} , Q_{θ_2} , and actor network π_{ϕ} with random parameters θ_1 , θ_2 , ϕ
- 2: Initialize target networks $\theta'_1 \leftarrow \theta_1$, $\theta'_2 \leftarrow \theta_2$, $\phi' \leftarrow \phi$
- 3: Initialize replay buffer B
- 4: **for** $t = 1$ to T **do**
- 5: Select action with exploration noise $a \sim \pi_{\phi}(s) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma)$ and observe reward R and new state s'
- 6: Store transition tuple (s, a, R, s') in B
- 7: Sample mini-batch of N transitions (s, a, R, s') from B
- 8: $\tilde{a} \leftarrow \pi_{\phi'}(s') + \epsilon$, $\epsilon \sim \text{clip}(\mathcal{N}(0, \tilde{\sigma}), -c, c)$
- 9: $y \leftarrow R + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \tilde{a})$
- 10: Update critics $\theta_i \leftarrow \arg \min_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(s, a))^2$
- 11: **if** $t \bmod d$ **then**
- 12: Update ϕ by the deterministic policy gradient:

$$\nabla_{\phi} J(\phi) = N^{-1} \sum \nabla_a Q_{\theta_1}(s, a) \Big|_{a=\pi_{\phi}(s)} \nabla_{\phi} \pi_{\phi}(s)$$
- 13: Update target networks:
- 14: $\theta'_i \leftarrow \tau \theta_i + (1 - \tau) \theta'_i$
- 15: $\phi' \leftarrow \tau \phi + (1 - \tau) \phi'$
- 16: **end if**
- 17: **end for**
- 18: **end for**

Συνολικά, οι τροποποιήσεις του αλγορίθμου TD3 είναι απλές και εύκολες στην υλοποίηση, ενώ παίζουν καθοριστικό ρόλο στην αύξηση της απόδοσης σε σχέση με τον αλγόριθμο DDPG. Συγκεκριμένα, οι αλλαγές του TD3 μειώνουν την αστάθεια που μπορεί να εμφανιστεί στον DDPG,

προωθούν την ανθεκτικότητα της εκπαίδευσης και τελικά, οδηγούν σε πιο αξιόπιστες πολιτικές.

Παρόλα αυτά, ο αλγόριθμος είναι πιο πολύπλοκος από τον DDPG, με χαρακτηριστικά παραδείγματα την χρήση δύο δικτύων κριτή και την καθυστέρηση της ενημέρωσης της πολιτικής. Επομένως, το υπολογιστικό κόστος του αλγορίθμου είναι υψηλότερο, ενώ η εκπαίδευση του μπορεί να απαιτεί περισσότερο χρόνο. Ακόμα, οι νέες τροποποιήσεις εισάγουν περισσότερες υπερ-παραμέτρους, που πρέπει να ρυθμιστούν με προσοχή, ώστε να μην επηρεάσουν αρνητικά την απόδοση του αλγορίθμου.

3.4.6 Ο αλγόριθμος SAC

Ο αλγόριθμος SAC (*Soft Actor-Critic*) είναι ένας αλγόριθμος βαθιάς ενισχυτικής μάθησης, που ανήκει στην κατηγορία των αλγορίθμων δράστη-κριτή. Αναπτύχθηκε το 2018 από ερευνητές του πανεπιστημιού του Berkeley και της Google (Haarnoja κ.ά. 2018). Αποτελεί μία επέκταση των παραδοσιακών μεθόδων δράστη-κριτή, που ενσωματώνει την κανονικοποίηση της εντροπίας στην αντικειμενική συνάρτηση, κάνοντας την πολιτική στοχαστική και ενθαρρύνοντας την εξερεύνηση. Εφαρμόζεται σε περιβάλλοντα με συνεχείς χώρους καταστάσεων και ενεργειών. Ξεχωρίζει για την αποδοτικότητα του στη χρήση δειγμάτων, την ανθεκτικότητα της εκπαίδευσης του και τις επιδόσεις του σε πολύπλοκα προβλήματα, όπως η πλοιήγηση ρομπότ.

Ο SAC βασίζεται στο πλαίσιο της μέγιστης εντροπίας της ενισχυτικής μάθησης. Σε αυτό το πλαίσιο, ο δράστης στοχεύει στη μεγιστοποίηση της αναμενόμενης ανταμοιβής, ενώ ταυτόχρονα μεγιστοποιεί την εντροπία. Με άλλα λόγια, ο δράστης προσπαθεί να επιτύχει στην εργασία του, ενεργώντας όσο το δυνατόν πιο τυχαία. Με τον τρόπο αυτό, ο αλγόριθμος επιχειρεί να πετυχεί μία ισορροπία στο δίλημμα εξερεύνησης-αξιοποίησης. Το πλαίσιο αυτό εμφανίζεται και στην αντικειμενική συνάρτηση του αλγορίθμου SAC, η οποία παρουσιάζεται στην εξίσωση 3.16:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_t R(s_t, a_t) - \alpha \log(\pi(a_t | s_t)) \right] \quad (3.16)$$

Οι όροι της αντικειμενικής συνάρτησης αναλύονται παρακάτω:

- Ο όρος $\sum_t R(s_t, a_t)$ αποτελεί την αναμενόμενη αθροιστική ανταμοιβή του περιβάλλοντος. Παρατηρούμε ότι ο πράκτορας στοχεύει στη μεγιστοποίηση της.
- Ο όρος $-\alpha \log(\pi(a_t | s_t))$ αποτελεί την εντροπία της πολιτικής, την οποία ο πράκτορας επίσης στοχεύει να μεγιστοποιήσει.
- Ο όρος α είναι η παράμετρος θερμοκρασίας (*temperature parameter*), η οποία καθορίζει την επίδραση του όρου εντροπίας έναντι του όρου της ανταμοιβής και συνεπώς, ελέγχει τη στοχαστικότητα της βέλτιστης πολιτικής. Η κλασική, μέγιστη αναμενόμενη ανταμοιβή, που χρησιμοποιείται στους συμβατικούς αλγορίθμους ενισχυτικής μάθησης, μπορεί να ανακτηθεί στο όριο όταν $\alpha \rightarrow 0$.

3.4 Βαθιά Ενισχυτική Μάθηση

Προκειμένου να μεγιστοποιήσει την παραπάνω αντικειμενική συνάρτηση, ο αλγόριθμος SAC εκπαιδεύει πέντε διαφορετικά νευρωνικά δίκτυα:

- 1 δίκτυο δράστη, το οποίο αντιπροσωπεύει την πολιτική του πράκτορα $\pi(a_t|s_t)$,
- 2 δίκτυα κριτή, τα οποία εκτιμούν την αξία των ζευγών κατάστασης-ενέργειας, δηλ. τις τιμές $Q(s_t, a_t)$ και
- 2 δίκτυα κριτή στόχου (*target networks*), τα οποία αποτελούν καθυστερημένα αντίγραφα των δικτύων κριτή και χρησιμοποιούνται στην εκπαίδευση των δικτύων κριτή, στον υπολογισμό του σφάλματος. τα δίκτυα αυτά συμβάλουν σημαντικά στη σταθεροποίηση της εκπαίδευσης.

Ακόμα, ο αλγόριθμος SAC εκτελείται off-policy, δηλ. τα δεδομένα που συλλέγονται από την αλληλεπίδραση με το περιβάλλον, αποθηκεύονται σε μία μνήμη (*replay buffer*) και μπορούν να χρησιμοποιηθούν πολλές φορές κατά την εκπαίδευση.

Αναλυτικότερα, η μεθοδολογία που ακολουθεί ο αλγορίθμος SAC δίνεται παράκατω υπό μορφή ψευδοκώδικα. Έχουν προστεθεί σχόλια στον ψευδοκώδικα, προκειμένου να γίνει πιο κατανοητή η λειτουργία του αλγορίθμου.

SAC Algorithm

```

1: Input:  $\theta_1, \theta_2, \phi$                                 ▷ Initial critic and actor networks parameters
2:  $\bar{\theta}_1 \leftarrow \theta_1, \bar{\theta}_2 \leftarrow \theta_2$     ▷ Initialize target network weights
3:  $D \leftarrow \emptyset$                                      ▷ Initialize an empty replay buffer
4: for each iteration do
5:   for each environment step do                         ▷ Sample action from the policy
6:      $a_t \sim \pi_\phi(a_t|s_t)$                             ▷ Receive transition from the environment
7:      $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$                       ▷ Store the transition in the replay buffer
8:      $D \leftarrow D \cup \{(s_t, a_t, R(s_t, a_t), s_{t+1})\}$ 
9:   end for
10:  for each gradient step do                           ▷ Update critic networks Q-functions
11:     $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$  for  $i \in \{1, 2\}$     ▷ Update actor network policy
12:     $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$                                ▷ Adjust temperature
13:     $\alpha \leftarrow \alpha - \lambda_\alpha \hat{\nabla}_\alpha J(\alpha)$                           ▷ Update target network weights
14:     $\bar{\theta}_i \leftarrow \tau \theta_i + (1 - \tau) \bar{\theta}_i$  for  $i \in \{1, 2\}$ 
15:  end for
16: end for
17: Output:  $\theta_1, \theta_2, \phi$                                 ▷ Optimized critic and actor networks parameters

```

Συνολικά, ο SAC διαθέτει πολλά στοιχεία στον τρόπο λειτουργίας του, τα οποία δημιουργούν σημαντικά πλεονεκτήματα σε σχέση με άλλους αλγορίθμους.

Αρχικά, η ιδιότητα του SAC ως off policy αλγορίθμου και η χρήση του replay buffer, τον καθιστά ιδιαίτερα αποδοτικό στη χρήση δειγμάτων και μειώνει τον αριθμό των απαιτούμενων αλληλεπιδράσεων με το περιβάλλον. Επιπλέον, η εκπαίδευση γίνεται με τη χρήση μιας στοχαστικής πολιτικής, η οποία ενθαρρύνει τον πράκτορα να εξερευνήσει το περιβάλλον.

Επίσης, όπως αναφέρθηκε και παραπάνω, η χρήση της εντροπίας στην αντικειμενική συνάρτηση, βοηθάει στην αποδοτική εξερεύνηση του περιβάλλοντος από τον πράκτορα και αποτρέπει την πρόωρη σύγκλιση της πολιτικής του. Αυτό είναι κρίσιμο σε περιβάλλοντα με αραιές ανταμοιβές, όπως το πρόβλημα της αυτόνομης στάθμευσης.

Ακόμα, η χρήση διαφορετικών νευρωνικών δικτύων για την εκτιμήση των τιμών Q , συμβάλλει στην αποφυγή της υπερεκτίμησης των τιμών τους, το οποίο αποτελεί συχνό πρόβλημα σε τέτοιου είδους μεθόδους. Έτσι, ο αλγόριθμος επιτυγχάνει πιο σταθερή και αξιόπιστη μάθηση.

Ωστόσο, ο αλγόριθμος SAC παρουσιάζει και ορισμένα μειονέκτημα. Το βασικότερο από αυτά αποτελεί η πολυπλοκότητα του, η οποία οδηγεί σε αυξημένο υπολογιστικό κόστος. Συγκεκριμένα, η ανάγκη ενημέρωσης πολλαπλών νευρωνικών δικτύων για τη συνάρτηση αξίας, επιβαρύνει σημαντικά το υπολογιστικό έργο.

Μία ακόμα αδυναμία του αλγορίθμου είναι η ευαισθησία στην κλίμακα των ανταμοιβών (*reward scale*). Συγκεκριμένα, για μικρές τιμές των ανταμοιβών, ο πράκτορας αποτυγχάνει να αξιοποιήσει το σήμα ανταμοιβής, με αποτέλεσμα σημαντική υποβάθμιση της απόδοσης του. Αντίθετα, για μεγάλες τιμές ανταμοιβών, ο πράκτορας μαθαίνει γρήγορα στην αρχή, αλλά συχνά συγκλίνει σε τοπικά ελάχιστα, λόγω έλλειψης επαρκούς εξερεύνησης. Επομένως, είναι απαραίτητη η σωστή κλιμάκωση των ανταμοιβών, ώστε ο πράκτορας να ισορροπήσει την εξερεύνηση και την αξιοποίηση και να πετύχει τελικά καλύτερη απόδοση. Παρόλα αυτά, είναι θετικό πως η κλίμακα των ανταμοιβών αποτελεί τη μόνη υπερ-παράμετρο, που απαιτεί στην πράξη, προσεκτική ρύθμιση από τον σχεδιαστή του συστήματος.

3.5 Σύνοψη

Στην Ενότητα αυτή, δώσαμε το θεωρητικό υπόβαθρο για την παρούσα εργασία. Ξεκινήσαμε ορίζοντας την Τεχνητή Νοημοσύνη και εξετάζοντας τις σύγχρονες εφαρμογές και τις κατηγορίες της. Στη συνέχεια, αναλύσαμε το πεδίο της Μηχανικής Μάθησης και μελετήσαμε τους τρεις τύπους της. Έπειτα, επικεντρώθηκαμε στο κύριο αντικείμενο της εργασίας, τον τομέα της Ενισχυτική Μάθηση. Ορίσαμε τις βασικές έννοιες του πεδίου και εξετάσαμε τη χρήση των Νευρωνικών Δικτύων σε αυτό. Τέλος, παρουσιάσαμε τις κατηγορίες αλγορίθμων Ενισχυτικής Μάθησης και αναλύσαμε τους 5 αλγορίθμους που υλοποιήσαμε στην παρούσα εργασία.

Η θεωρία που αναλύθηκε σε αυτήν την Ενότητα αποτελεί τη βάση για την κατανόηση των εκπαιδεύσεων που διεξήχθησαν και των αποτελεσμάτων τους, τα οποία θα δούμε στις επομένες Ενότητες. Επομένως, από την επόμενη Ενότητα, αρχίζει το πρακτικό κομμάτι της εργασίας, ξεκινώντας με την παρουσιάση του περιβάλλοντος εκπαίδευσης, δηλ. του παιχνιδιού που κατασκευάστηκε.

Βιβλιογραφία

- AI-Forge. 2022. ‘A.I. Learns To Drive –youtube.com’ . https://www.youtube.com/watch?v=ZshliCIM9ZA&ab_channel=AIForge.
- Alvarez, Waldo. 2017. ‘Markov decision process - Wikipedia –en.wikipedia.org’ . https://en.wikipedia.org/wiki/Markov_decision_process.
- Anzalone, Luca, Paola Barra, Silvio Barra, Aniello Castiglione, και Michele Nappi. 2022. ‘An End-to-End Curriculum Learning Approach for Autonomous Driving Scenarios’. *IEEE Transactions on Intelligent Transportation Systems* 23 (10): 19817–26. <https://doi.org/10.1109/TITS.2022.3160673>.
- Arzt, Samuel. 2019. ‘AI Learns to Park - Deep Reinforcement Learning –youtube.com’ . https://www.youtube.com/watch?v=VMp6pq6_QjI&t=462s&ab_channel=SamuelArzt.
- Atomwise. 2018. ‘AtomNet® Technology has the Power to Impact Early Drug Discovery – blog.atomwise.com’ . <https://blog.atomwise.com/atomnet-technology-has-the-power-to-impact-early-drug-discovery>.
- Baeldung. 2023. ‘Epsilon-Greedy Q-learning –baeldung.com’ . <https://www.baeldung.com/cs/epsilon-greedy-q-learning>.
- Blanco, Sebastian. 2023. ‘Report: Tesla Autopilot Involved in 736 Crashes since 2019 –caranddriver.com’ . <https://www.caranddriver.com/news/a44185487/report-tesla-autopilot-crashes-since-2019/>; CarandDriver.com.
- BostonDynamics. 2024. ‘Starting on the Right Foot with Reinforcement Learning | Boston Dynamics –bostondynamics.com’ . <https://bostondynamics.com/blog/starting-on-the-right-foot-with-reinforcement-learning/>.
- Bright Side. 2020. ‘Self-Driving Cars: 7 Pros and 7 Cons –youtube.com’ . https://www.youtube.com/watch?v=9RAFgKcY4uA&ab_channel=BRIGHTSIDE.
- Code-Bullet. 2019. ‘A.I. Learns to DRIVE –youtube.com’ . https://www.youtube.com/watch?v=r428O_CMcpI&ab_channel=CodeBullet.
- Comsol. 2023. ‘Deep Neural Network –doc.comsol.com’ . https://doc.comsol.com/6.2/doc/com.comsol.help.comsol/comsol_ref_definitions.19.050.html.
- DeepMind. 2016. ‘AlphaGo –deepmind.google’ . <https://deepmind.google/technologies/alphago/>.

- Deichmann, Johannes, Eike Ebel, Kersten Heineke, Ruth Heuss, Martin Kellner, και Fabian Steiner. 2023. ‘Autonomous driving’ s future: Convenient and connected –mckinsey.com’ . <https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/autonomous-drivings-future-convenient-and-connected>; McKinsey.
- EITCA. 2024. ‘Does increasing of the number of neurons in an artificial neural network layer increase the risk of memorization leading to overfitting? - EITCA Academy –eitca.org’ . <https://eitca.org/artificial-intelligence/eitca-ai-tff-tensorflow-fundamentals/overfitting-and-underfitting-problems/solving-models-overfitting-and-underfitting-problems-part-1/does-increasing-of-the-number-of-neurons-in-an-artificial-neural-network-layer-increase-the-risk-of-memorization-leading-to-overfitting/>.
- Erickson, Jim. 2018. ‘Maximizing the environmental benefits of autonomous vehicles –news.umich.edu’ . <https://news.umich.edu/maximizing-the-environmental-benefits-of-autonomous-vehicles/>; University of Michigan.
- Ezeokeke, Emmanuel. 2024. ‘AI Agents vs. AI Models: Why Agents Take the Lead –linkedin.com’ . <https://www.linkedin.com/pulse/ai-agents-vs-models-why-take-lead-emmanuel-ezeokeke-gv4xf/>; LinkedIn.
- Fujimoto, Scott, Herke Hoof, και David Meger. 2018. ‘Addressing Function Approximation Error in Actor-Critic Methods’ . <https://arxiv.org/abs/1802.09477>.
- Gillis, Alexander. 2024. ‘What is a self-driving car? | Definition from TechTarget –techttarget.com’ . <https://www.techttarget.com/searchenterpriseai/definition/driverless-car>; TechTarget.
- Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, και Sergey Levine. 2018. ‘Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor’ . <https://arxiv.org/abs/1801.01290>.
- Hanson-Robotics. 2016. ‘Sophia - Hanson Robotics –hansonrobotics.com’ . <https://www.hansonrobotics.com/sophia/>.
- Johnson, Jonathan. 2020. ‘What’s a Deep Neural Network? Deep Nets Explained –bmc.com’ . <https://www.bmc.com/blogs/deep-neural-network/>; BMC Software.
- Karpathy, Andrej. 2016. ‘Neural Networks Part 1: Setting Up the Architecture. Notes for CS231n Convolutional Neural Networks for Visual Recognition’ . Stanford University: Stanford, CA, USA.
- Lai, Leonardo. 2018. ‘Automatic Parking with Q-Learning’ . GitHub repository. <https://github.com/leol2/Autoparking>; GitHub. <https://doi.org/10.5281/zenodo.4568892>.
- Lateef, Zulaikha. 2024. ‘Types of AI: Understanding Different Types of Artificial Intelligence in 2024 –edureka.co’ . <https://www.edureka.co/blog/types-of-artificial-intelligence/>; Edureka.

- Lee, Dan. 2019. ‘Reinforcement Learning, Part 1: A Brief Introduction —medium.com’ . <https://medium.com/ai%C2%B3-theory-practice-business/reinforcement-learning-part-1-a-brief-introduction-a53a849771cf>.
- Lillicrap, Timothy, Jonathan Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, και Daan Wierstra. 2015. ‘Continuous control with deep reinforcement learning’ . <https://arxiv.org/abs/1509.02971>.
- Loiacono, Daniele, Alessandro Prete, Pier Luca Lanzi, και Luigi Cardamone. 2010. ‘Learning to overtake in TORCS using simple reinforcement learning’ . Στο *IEEE Congress on Evolutionary Computation*, 1–8. <https://doi.org/10.1109/CEC.2010.5586191>.
- McCarthy, John. 2004. ‘What is Artificial Intelligence?’ , Ιανουάριος.
- McMurray, Alex. 2023. ‘Goldman Sachs AI head likes technique that ’ creates a trader’ —efinancialcareers.com’ . <https://www.efinancialcareers.com/news/2023/10/will-ai-replace-traders-goldman-sachs>.
- Melo, Francisco. 2001. ‘Convergence of Q-learning: a simple proof’ . <http://users.isr.ist.utl.pt/~mtjspan/readingGroup/ProofQlearning.pdf>.
- Metz, Cade. 2016. ‘In Two Moves, AlphaGo and Lee Sedol Redefined the Future —wired.com’ . <https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/>.
- Mitchell, Thomas. 1997. *Machine Learning*. McGraw-Hill series in computer science. New York, NY: McGraw-Hill Professional.
- Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, κ.ά. 2015. ‘Human-level control through deep reinforcement learning’ . *Nature* 518: 529–33. <https://api.semanticscholar.org/CorpusID:205242740>.
- Moreira, Dinis. 2021. ‘Deep Reinforcement Learning for Automated Parking’ . <https://repositorio-aberto.up.pt/handle/10216/136074>; University of Porto.
- Mujumdar, Pranav, Parthiv Shah, και Xinyue Cui. 2020. ‘Autonomous Car Parking Simulator using Unity MLAgents’ . <https://github.com/pranavmujumdar/CarParkingAi>; Northeastern University.
- NeuralNine. 2021. ‘Self-Driving AI Car Simulation in Python —youtube.com’ . https://www.youtube.com/watch?v=Cy155O5R1Oo&ab_channel=NeuralNine.
- OpenAI. 2018. ‘Part 2: Kinds of RL Algorithms — Spinning Up documentation —spinningup.openai.com’ . https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html.
- OpenAI. 2019. ‘OpenAI Five defeats Dota 2 world champions’ . <https://openai.com/index/openai-five-defeats-dota-2-world-champions/>.

- Papers With Code. 2024. ‘Papers with Code - PPO Explained —paperswithcode.com’ . <https://paperswithcode.com/method/ppo>.
- Parkinson, Avery. 2019. ‘The Epsilon-Greedy Algorithm for Reinforcement Learning —medium.com’ . <https://medium.com/analytics-vidhya/the-epsilon-greedy-algorithm-for-reinforcement-learning-5fe6f96dc870>; Analytics Vidhya.
- Porro, Marco. 2020. ‘Self Driving Car Neural Network - with Python and NEAT (CODE in the description) —youtube.com’ . https://www.youtube.com/watch?v=cFjYinc465M&ab_channel=MarcoPorro.
- Prijono, Benny. 2020. ‘GitHub - bennylp/RL-Taxonomy: Loose taxonomy of reinforcement learning algorithms —github.com’ . <https://github.com/bennylp/RL-Taxonomy>.
- Puigdomenech, Adria, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, και Charles Blundell. 2020. ‘Agent57: Outperforming the human Atari benchmark —deepmind.google’ . <https://deepmind.google/discover/blog/agent57-outperforming-the-human-atari-benchmark/>.
- Quang, Luu. 2024. ‘Q-Learning vs. Deep Q-Learning vs. Deep Q-Network | Baeldung on Computer Science —baeldung.com’ . https://www.baeldung.com/cs/q-learning-vs-deep-q-learning-vs-deep-q-network?fbclid=IwZKh0bgNhZW0CMTEAAR2SFVuY7XwXuXgcuva0kKQ24Y_xpBt4A6Ajsq79KfHwmyvdLOsOQD1MrmM_aem_6t-0YIbksWbZFNA1XWkNhw.
- Russell, Stuart, και Peter Norvig. 2021. *Artificial intelligence: A modern approach, global edition*. 4ο έκδ. London, England: Pearson Education.
- Saini, Sajan. 2019. ‘How do self-driving cars 'see'? - Sajan Saini —youtube.com’ . https://www.youtube.com/watch?v=PRg5RNU_JLk&ab_channel=TED-Ed.
- Samuel, A. L. 1959. ‘Some studies in machine learning using the game of checkers’ . *IBM Journal of Research and Development* 44 (1.2): 206–26. <https://doi.org/10.1147/rd.441.0206>.
- Saravia, Felipe. 2020. ‘Neural Network Cars and Genetic Algorithms (1/2) —youtube.com’ . https://www.youtube.com/watch?v=-sg-GgoFCP0&ab_channel=FelipeSaravia.
- Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, και Oleg Klimov. 2017. ‘Proximal Policy Optimization Algorithms’ . <https://arxiv.org/abs/1707.06347>.
- Singh, S. 2015. ‘Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey’ . <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115>; NHTSA’s National Center for Statistics; Analysis.
- Society of Automobile Engineers. 2021. ‘SAE Levels of Driving Automation™ Refined for Clarity and International Audience —sae.org’ . <https://www.sae.org/blog/sae-j3016-update>.

- Stewart, Ellis. 2023. ‘What is Machine Learning (ML)? Types, Models, Algorithms | Enterprise Tech News EM360 —em360tech.com’. <https://em360tech.com/tech-article/what-is-machine-learning-ml>; EM360 Tech.
- Sutton, Richard S, και Andrew G Barto. 2018. *Reinforcement learning: An introduction*. 2o έκδ. Cambridge, MA: MIT press.
- Szymanski, Lech. 2018. ‘Course COSC470’ . <https://www.cs.otago.ac.nz/cosc470/>; University of Otago, Computer Science Department.
- Tesauro, Gerald. 1994. ‘TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play’ . *Neural Computation* 6 (2): 215–19. <https://doi.org/10.1162/neco.1994.6.2.215>.
- Thomson, T, και Ryan Thomas. 2023. ‘Ageism, sexism, classism and more: 7 examples of bias in AI-generated images –theconversation.com’ . <https://theconversation.com/ageism-sexism-classism-and-more-7-examples-of-bias-in-ai-generated-images-208748#:~:text=There%20were%20also%20notable%20differences,of%20more%20fluid%20gender%20expression>.
- Trudell, Craig. 2024. ‘Bloomberg - Are you a robot? —bloomberg.com’ . <https://www.bloomberg.com/news/newsletters/2024-04-25/elon-musk-s-tesla-robotaxi-predictions-were-all-wrong>; Bloomberg.
- Tucker, Sean. 2024. ‘Self-Driving Cars: Everything You Need To Know - Kelley Blue Book —kbb.com’ . <https://www.kbb.com/car-advice/self-driving-cars/>; Kelley Blue Book.
- Turing, A. M. 1950. ‘I.—COMPUTING MACHINERY AND INTELLIGENCE’ . *Mind* LIX (236): 433–60. <https://doi.org/10.1093/mind/LIX.236.433>.
- University of Michigan Center for Sustainable Systems. 2023. ‘Autonomous Vehicles Factsheet —css.umich.edu’ . <https://css.umich.edu/publications/factsheets/mobility/autonomous-vehicles-factsheet>.
- Watkins, Christopher. 1989. ‘Learning From Delayed Rewards’ , Ιανουάριος.
- Zand, Aria, Zack Stokes, Arjun Sharma, Welmoed van Deen, και Daniel Hommes. 2022. ‘Artificial Intelligence for Inflammatory Bowel Diseases (IBD); Accurately Predicting Adverse Outcomes Using Machine Learning’ . *Digestive Diseases and Sciences* 67 (Απρίλιος): 1–12. <https://doi.org/10.1007/s10620-022-07506-8>.

