Σχολή Θετικών Επιστημών και Τεχνολογίας

Πρόγραμμα Σπουδών Πληροφορική

Πτυχιακή Εργασία

Ανάπτυξη συστήματος αυτόματης ενδοημερήσιας διαπραγμάτευσης με αρχιτεκτονική Dueling Deep Q-Network και ενσωμάτωση του δείκτη VWAP

Αθανάσιος Χαμπέος

Επιβλέπων καθηγητής: Δρ. Ευστράτιος Γεωργόπουλος

Πάτρα, Ιούλιος 2026

Ανάπτυξη συστήματος αυτόματης ενδοημερήσιας διαπραγμάτευσης με αρχιτεκτονική Dueling Deep Q-Network και ενσωμάτωση του δείκτη VWAP

Αθανάσιος Χαμπέος

Επιτροπή Επίβλεψης Πτυχιακής / Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής:
Δρ. Ευστράτιος Γεωργόπουλος
Τμήμα Διοίκησης Επιχειρήσεων
Επιστήμης & Οργανισμών
Πανεπιστημίου Πελοποννήσου
Συνεργαζόμενο Εκπαιδευτικό
Προσωπικό (ΣΕΠ)
Ελληνικό Ανοιχτό Πανεπιστήμιο
(ΕΑΠ)

Συν-Επιβλέπων Καθηγητής:
Δρ. Πλαγιανάκος Βασίλειος
Τμήμα Πληροφορικής με
Εφαρμογές στη Βιοϊατρική
Πανεπιστήμιο Θεσσαλίας

Συν-Επιβλέπων Καθηγητής:
Δρ. Αναγνωστόπουλος Χρήστος –
Νικόλαος
Τμήμα Πολιτισμικής Τεχνολογίας
και Επικοινωνίας
Πανεπιστήμιο Αιγαίου

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τον Καθηγητή Δρ. Ευστράτιο Γεωργόπουλο για την πολύτιμη καθοδήγηση, την επιστημονική του συμβολή και τη συνεχή υποστήριξή του καθ' όλη τη διάρκεια εκπόνησης της παρούσας εργασίας.

Επιπλέον, θα ήθελα να ευχαριστήσω θερμά τον Δρ. Θωμά Αμοργιανιώτη για τη σημαντική βοήθεια, τις εύστοχες παρατηρήσεις και τη διαρκή ενθάρρυνσή του, οι οποίες συνέβαλαν ουσιαστικά στην ολοκλήρωση της εργασίας.

Περίληψη

Η τεχνητή νοημοσύνη αποτελεί βασικό παράγοντα καινοτομίας στον τομέα των χρηματοοικονομικών, επηρεάζοντας σημαντικά τον τρόπο με τον οποίο αναλύονται τα δεδομένα της αγοράς και λαμβάνονται επενδυτικές αποφάσεις. Ιδιαίτερα, οι μέθοδοι ενισχυτικής μάθησης προσφέρουν ένα δυναμικό πλαίσιο για την ανάπτυξη συστημάτων που μπορούν να προσαρμόζονται σε μεταβαλλόμενες συνθήκες και να βελτιστοποιούν τη στρατηγική τους μέσω εμπειρίας.

Η παρούσα εργασία διερευνά την εφαρμογή της ενισχυτικής μάθησης στο πρόβλημα της αυτόματης διαπραγμάτευσης μετοχών, αντιμετωπίζοντας τη διαδικασία ως πρόβλημα διαδοχικής λήψης αποφάσεων. Στο πλαίσιο αυτό, αναπτύσσεται ένας πράκτορας βασισμένος στην αρχιτεκτονική Dueling Deep Q-Network (Dueling DQN), ο οποίος εκπαιδεύεται σε προσομοιωμένο περιβάλλον ενδοημερήσιας διαπραγμάτευσης.

Η μεθοδολογία περιλαμβάνει τον σχεδιασμό περιβάλλοντος που αναπαριστά τη δυναμική της αγοράς μέσω ιστορικών δεδομένων τιμών και όγκου συναλλαγών, καθώς και τον ορισμό συνάρτησης ανταμοιβής βασισμένης στη μεταβολή της αξίας του χαρτοφυλακίου (Profit and Loss – PnL). Ιδιαίτερη έμφαση δίνεται στην ενσωμάτωση του δείκτη Volume-Weighted Average Price (VWAP) ως πρόσθετου χαρακτηριστικού στο διάνυσμα κατάστασης του πράκτορα, με στόχο τη διερεύνηση της επίδρασής του στη διαδικασία λήψης αποφάσεων.

Η αξιολόγηση της προτεινόμενης προσέγγισης πραγματοποιείται μέσω δύο διακριτών πειραματικών σεναρίων και συγκριτικών πειραμάτων μεταξύ των εκδοχών BASE και VWAP, με στόχο τη διερεύνηση της επίδρασης της ενσωμάτωσης του δείκτη VWAP στη στρατηγική συμπεριφορά και στη συνολική απόδοση του πράκτορα. Τα αποτελέσματα δείχνουν ότι ο πράκτορας είναι ικανός να αναπτύσσει αποτελεσματικές στρατηγικές ενδοημερήσιας διαπραγμάτευσης, ενώ η ενσωμάτωση των χαρακτηριστικών VWAP διαφοροποιεί τη διαδικασία λήψης αποφάσεων χωρίς να οδηγεί συστηματικά σε υψηλότερη τελική απόδοση.

Λέξεις – Κλειδιά

Τεχνητή Νοημοσύνη, Ενισχυτική Μάθηση, Deep Reinforcement Learning, Dueling Deep Q-Network, Αλγοριθμική Διαπραγμάτευση, VWAP

Abstract

Artificial Intelligence has become a key driver of innovation in the financial sector, significantly influencing the way market data are analyzed and investment decisions are made. In particular, reinforcement learning provides a powerful framework for developing adaptive systems that improve their performance through interaction with dynamic environments.

This study investigates the application of reinforcement learning to the problem of automated stock trading, treating it as a sequential decision-making process under uncertainty. In this context, a trading agent based on the Dueling Deep Q-Network (Dueling DQN) architecture is developed and trained within a simulated intraday trading environment.

The proposed methodology includes the design of a trading environment that models market dynamics using historical price and volume data, as well as the definition of a reward function based on changes in portfolio value (Profit and Loss – PnL). Special emphasis is placed on the integration of the Volume-Weighted Average Price (VWAP) as an additional feature in the agent's state representation, aiming to investigate its impact on the decision-making process.

The proposed approach is evaluated through two experimental scenarios and comparative experiments between the BASE and VWAP configurations, investigating the effect of including VWAP as an additional state feature on both the agent's trading behaviour and overall performance. The results show that the Dueling DQN agent is capable of learning effective intraday trading strategies, while the inclusion of VWAP-derived features influences the agent's decision-making behaviour without consistently improving the final trading performance.

Keywords

Artificial Intelligence, Reinforcement Learning, Deep Reinforcement Learning, Dueling Deep Q-Network, Algorithmic Trading, VWAP

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Ψευδοκωδικών

Συντομογραφίες & Ακρωνύμια

DQN — Deep Q-Network

DRL — Deep Reinforcement Learning

Dueling DQN — Dueling Deep Q-Network

MACD — Moving Average Convergence Divergence

MDP — Markov Decision Process

PnL — Profit and Loss

RL — Reinforcement Learning

RSI — Relative Strength Index

VWAP — Volume Weighted Average Price

1. Εισαγωγή

Η ραγδαία εξέλιξη των χρηματοοικονομικών αγορών και η αυξανόμενη διαθεσιμότητα δεδομένων υψηλής συχνότητας έχουν οδηγήσει στη διεύρυνση της χρήσης αυτοματοποιημένων συστημάτων διαπραγμάτευσης. Η αλγοριθμική διαπραγμάτευση αποτελεί πλέον βασικό μηχανισμό λειτουργίας των σύγχρονων αγορών, καθώς επιτρέπει την ταχεία εκτέλεση συναλλαγών, τη μείωση του ανθρώπινου παράγοντα και την αξιοποίηση σύνθετων υπολογιστικών μεθόδων για τη λήψη επενδυτικών αποφάσεων.

Παρά την πρόοδο αυτή, η ανάπτυξη αποδοτικών και ρεαλιστικών συστημάτων αυτόματης διαπραγμάτευσης παραμένει ένα σύνθετο πρόβλημα. Οι παραδοσιακές προσεγγίσεις βασίζονται είτε σε στατικούς κανόνες τεχνικής ανάλυσης είτε σε προβλεπτικά μοντέλα, τα οποία συχνά αδυνατούν να αντιμετωπίσουν τη δυναμική και στοχαστική φύση των αγορών. Στο πλαίσιο αυτό, οι μέθοδοι ενισχυτικής μάθησης έχουν αναδειχθεί ως ιδιαίτερα κατάλληλες, καθώς προσεγγίζουν τη διαπραγμάτευση ως πρόβλημα διαδοχικής λήψης αποφάσεων.

Η παρούσα εργασία εστιάζει στην ανάπτυξη και αξιολόγηση ενός συστήματος αυτόματης ενδοημερήσιας διαπραγμάτευσης που αξιοποιεί τεχνικές βαθιάς ενισχυτικής μάθησης, με έμφαση στη διερεύνηση της επίδρασης της ενσωμάτωσης πρόσθετων τεχνικών χαρακτηριστικών αγοράς, όπως ο δείκτης Volume-Weighted Average Price (VWAP), στο διάνυσμα κατάστασης του πράκτορα, με στόχο τη βελτίωση της διαδικασίας λήψης αποφάσεων και της συνολικής απόδοσης.

Η αξιολόγηση της προτεινόμενης προσέγγισης πραγματοποιείται σε δύο διακριτά πειραματικά σενάρια, τα οποία διαφοροποιούνται ως προς τη χρονική έκταση των δεδομένων και την κλίμακα της πειραματικής διαδικασίας. Το πρώτο αποτελεί ένα πιλοτικό σενάριο έξι διαδοχικών χρηματιστηριακών συνεδριάσεων, το οποίο χρησιμοποιείται για την αρχική επαλήθευση της λειτουργίας του συστήματος, ενώ το δεύτερο εφαρμόζει την ίδια μεθοδολογία walk-forward σε σύνολο 111 συνεδριάσεων, επιτρέποντας την αξιολόγηση της γενίκευσης και της σταθερότητας της συμπεριφοράς του πράκτορα σε σημαντικά μεγαλύτερο χρονικό ορίζοντα.

1.1 Αντικείμενο και σκοπός της εργασίας

Αντικείμενο της παρούσας εργασίας είναι η μελέτη και υλοποίηση ενός συστήματος αυτόματης διαπραγμάτευσης βασισμένου σε αρχιτεκτονική Dueling Deep Q-Network, το οποίο λαμβάνει αποφάσεις αγοράς, πώλησης ή αναμονής σε ενδοημερήσιο χρονικό ορίζοντα. Το σύστημα εκπαιδεύεται σε προσομοιωμένο περιβάλλον χρηματοοικονομικής αγοράς και αξιολογείται ως προς την αποδοτικότητα και τη σταθερότητά του.

Κύριος σκοπός της εργασίας είναι η διερεύνηση του κατά πόσο η ενσωμάτωση του δείκτη Volume-Weighted Average Price (VWAP) ως πρόσθετου χαρακτηριστικού πληροφορίας στο διάνυσμα κατάστασης ενός πράκτορα ενισχυτικής μάθησης μπορεί να επηρεάσει τη διαδικασία λήψης αποφάσεων, τη στρατηγική συμπεριφορά και τη συνολική απόδοση του συστήματος, σε σύγκριση με μια αντίστοιχη υλοποίηση χωρίς τα χαρακτηριστικά αυτά.

1.2 Ορισμός και περιγραφή του προβλήματος

Το πρόβλημα που εξετάζεται στην παρούσα εργασία αφορά τη λήψη διαδοχικών αποφάσεων διαπραγμάτευσης σε ένα δυναμικό και στοχαστικό περιβάλλον αγοράς. Σε κάθε χρονική στιγμή, το σύστημα καλείται να αποφασίσει αν θα προβεί σε αγορά, πώληση ή αναμονή, λαμβάνοντας υπόψη τη διαθέσιμη πληροφορία.

Η δυσκολία του προβλήματος δεν περιορίζεται μόνο στην πρόβλεψη της μελλοντικής πορείας των τιμών, αλλά κυρίως στη διαμόρφωση μιας αποδοτικής στρατηγικής διαδοχικών αποφάσεων σε συνθήκες αβεβαιότητας και υψηλής μεταβλητότητας. Το σύστημα καλείται να αξιοποιήσει πολυπαραγοντική πληροφορία της αγοράς, συμπεριλαμβανομένων τεχνικών δεικτών όπως ο VWAP, να διαμορφώσει στρατηγικές συναλλαγών και να επιδιώξει τη μεγιστοποίηση της συνολικής απόδοσης του χαρτοφυλακίου μέσω της αλληλεπίδρασής του με το περιβάλλον. Συνεπώς, το πρόβλημα ορίζεται ως πρόβλημα διαδοχικής λήψης αποφάσεων με στόχο τη μεγιστοποίηση της μακροπρόθεσμης κερδοφορίας.

1.3 Σημασία και ερευνητικό κίνητρο

Η σημασία του εξεταζόμενου προβλήματος είναι τόσο θεωρητική όσο και πρακτική. Σε θεωρητικό επίπεδο, η εργασία συμβάλλει στη διερεύνηση του τρόπου με τον οποίο πρόσθετοι τεχνικοί δείκτες αγοράς μπορούν να ενσωματωθούν ενεργά στη διαδικασία μάθησης συστημάτων ενισχυτικής μάθησης, ενισχύοντας την πληροφοριακή αναπαράσταση της κατάστασης του πράκτορα. Σε πρακτικό επίπεδο, η αξιοποίηση εμπλουτισμένων δεδομένων αγοράς μπορεί να βελτιώσει τη λήψη αποφάσεων σε συστήματα αυτόματης διαπραγμάτευσης, ιδιαίτερα σε περιβάλλοντα υψηλής μεταβλητότητας και ταχείας χρονικής εξέλιξης.

Το ερευνητικό κίνητρο προκύπτει από το γεγονός ότι, παρά την ευρεία χρήση του δείκτη VWAP στις χρηματοοικονομικές αγορές, η πλειονότητα των σχετικών προσεγγίσεων τον αξιοποιεί κυρίως ως εργαλείο ανάλυσης ή benchmark και όχι ως ενσωματωμένο χαρακτηριστικό στη μαθησιακή διαδικασία πρακτόρων ενισχυτικής μάθησης. Η εργασία επιδιώκει να διερευνήσει κατά πόσο η ενσωμάτωση του VWAP ως πρόσθετου χαρακτηριστικού εισόδου επηρεάζει τη διαδικασία λήψης αποφάσεων, τη στρατηγική συμπεριφορά και τη συνολική απόδοση ενός πράκτορα ενισχυτικής μάθησης, σε σύγκριση με μια αντίστοιχη υλοποίηση χωρίς τα χαρακτηριστικά αυτά.

1.4 Στόχοι και ερευνητικά ερωτήματα

Με βάση τα παραπάνω, οι βασικοί στόχοι της παρούσας εργασίας είναι:

- η ανάπτυξη ενός συστήματος αυτόματης διαπραγμάτευσης βασισμένου σε Dueling Deep Q-Network,
- η ενσωμάτωση του δείκτη VWAP ως πρόσθετου χαρακτηριστικού στο διάνυσμα κατάστασης του πράκτορα,
- η αξιολόγηση της επίδρασης της ενσωμάτωσης πρόσθετης πληροφορίας αγοράς στη στρατηγική συμπεριφορά και στη συνολική απόδοση του πράκτορα, και
- η διερεύνηση της σταθερότητας της στρατηγικής και της συμπεριφοράς του πράκτορα σε διαφορετικές συνθήκες αγοράς.

Τα αντίστοιχα ερευνητικά ερωτήματα που τίθενται είναι:

- μπορεί ένας πράκτορας ενισχυτικής μάθησης να μάθει αποτελεσματικές στρατηγικές ενδοημερήσιας διαπραγμάτευσης βασισμένος σε ιστορικά δεδομένα αγοράς;
- πώς επηρεάζει η ενσωμάτωση του δείκτη VWAP στο διάνυσμα κατάστασης τη στρατηγική συμπεριφορά και την απόδοση του μοντέλου σε σχέση με τη βασική εκδοχή (BASE);
- παραμένει σταθερή η στρατηγική του πράκτορα όταν η ίδια μεθοδολογία εφαρμόζεται σε πειραματικά σενάρια διαφορετικής χρονικής κλίμακας;

1.5 Δομή της Πτυχιακής Εργασίας

Η εργασία οργανώνεται ως εξής:

Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο, καλύπτοντας την εξέλιξη της αλγοριθμικής διαπραγμάτευσης, τις βασικές μεθόδους ανάλυσης χρηματοοικονομικών δεδομένων και τις σύγχρονες προσεγγίσεις μηχανικής και ενισχυτικής μάθησης. Ιδιαίτερη έμφαση δίνεται στους αλγορίθμους Deep Q-Networks, στην αρχιτεκτονική Dueling Deep Q-Network και στη θεωρητική αξιοποίηση του δείκτη Volume-Weighted Average Price (VWAP) ως πρόσθετου τεχνικού χαρακτηριστικού αγοράς.

Στο Κεφάλαιο 3 περιγράφεται αναλυτικά η μεθοδολογία της εργασίας, συμπεριλαμβανομένης της αρχιτεκτονικής του προτεινόμενου συστήματος, της συνάρτησης ανταμοιβής, της ενσωμάτωσης του VWAP στο διάνυσμα κατάστασης του πράκτορα, της προεπεξεργασίας των ενδοημερήσιων δεδομένων ανάλυσης ανά λεπτό, καθώς και του συνολικού πλαισίου προσομοίωσης, εκπαίδευσης και αξιολόγησης. Παρουσιάζονται επίσης τα δύο βασικά πειραματικά σενάρια της εργασίας, τα οποία διαφοροποιούνται ως προς τη χρονική έκταση των δεδομένων και την κλίμακα της διαδικασίας walk-forward εκτέλεσης.

Στο Κεφάλαιο 4 παρουσιάζονται η αρχιτεκτονική, η τεχνική υλοποίηση, οι βασικές διαδικασίες λειτουργίας, οι ψευδοκώδικες, καθώς και η εφαρμογή των δύο πειραματικών σεναρίων του συστήματος, με έμφαση στις διαφοροποιήσεις μεταξύ των εκδοχών BASE και VWAP.

Τέλος, στο Κεφάλαιο 5 παρουσιάζονται και αναλύονται τα αποτελέσματα των δύο πειραματικών σεναρίων, πραγματοποιείται συγκριτική αξιολόγηση μεταξύ των εκδοχών BASE και VWAP, εξετάζεται η συμπεριφορά και η σταθερότητα του πράκτορα σε διαφορετικές συνθήκες αγοράς και συνοψίζονται τα βασικά συμπεράσματα της εργασίας, ενώ παράλληλα προτείνονται κατευθύνσεις για μελλοντική έρευνα.

2. Θεωρητικό υπόβαθρο

2.1 Η εξέλιξη της αλγοριθμικής διαπραγμάτευσης

Η αλγοριθμική διαπραγμάτευση αποτελεί προϊόν της μακροχρόνιας εξέλιξης των χρηματοοικονομικών αγορών και της σταδιακής ενσωμάτωσης της πληροφορικής στη διαδικασία λήψης επενδυτικών αποφάσεων. Στα πρώτα στάδια λειτουργίας των αγορών, οι συναλλαγές βασίζονταν κυρίως στην ανθρώπινη κρίση, στην εμπειρία των επενδυτών και στη θεμελιώδη ανάλυση οικονομικών δεδομένων. Η διαδικασία αυτή χαρακτηριζόταν από περιορισμένη ταχύτητα, αυξημένο λειτουργικό κόστος και έντονη επίδραση ψυχολογικών παραγόντων.

Με την πρόοδο των υπολογιστικών συστημάτων και τη διάδοση των ηλεκτρονικών αγορών, εμφανίστηκαν οι πρώτες μορφές αυτοματοποιημένων στρατηγικών βασισμένων σε μαθηματικά και στατιστικά κριτήρια. Οι στρατηγικές αυτές επέτρεψαν την αυτοματοποίηση επιμέρους αποφάσεων, όπως η ενεργοποίηση εντολών αγοράς ή πώλησης βάσει προκαθορισμένων κανόνων, μειώνοντας τον χρόνο αντίδρασης και ενισχύοντας τη συνέπεια εφαρμογής των επενδυτικών στρατηγικών. Σε αυτή τη φάση, η τεχνική ανάλυση αποτέλεσε βασικό πυλώνα ανάπτυξης αλγοριθμικών συστημάτων, αξιοποιώντας δείκτες που εξάγονταν από ιστορικά δεδομένα τιμών και όγκου συναλλαγών (Murphy, 1999).

Η ραγδαία αύξηση της υπολογιστικής ισχύος, η διαθεσιμότητα δεδομένων υψηλής συχνότητας και η ανάπτυξη προηγμένων ηλεκτρονικών πλατφορμών για τη διενέργεια συναλλαγών οδήγησαν στη σταδιακή καθιέρωση της αλγοριθμικής διαπραγμάτευσης ως κυρίαρχου μηχανισμού λειτουργίας των σύγχρονων αγορών. Τα συστήματα αυτά είναι σε θέση να αναλύουν μεγάλους όγκους δεδομένων σε πραγματικό χρόνο, να εκτελούν συναλλαγές με ελάχιστη καθυστέρηση και να λειτουργούν χωρίς άμεση ανθρώπινη παρέμβαση, περιορίζοντας σημαντικά την επίδραση συναισθηματικών παραγόντων στη λήψη αποφάσεων.

Στη σύγχρονη βιβλιογραφία, η αλγοριθμική διαπραγμάτευση συνδέεται ολοένα και περισσότερο με τεχνικές τεχνητής νοημοσύνης και μηχανικής μάθησης, οι οποίες επιτρέπουν τη μετάβαση από στατικές στρατηγικές σε προσαρμοστικά συστήματα που μαθαίνουν από τα δεδομένα και την εμπειρία. Ιδιαίτερα, οι μέθοδοι ενισχυτικής μάθησης αντιμετωπίζουν τη διαπραγμάτευση ως πρόβλημα διαδοχικών αποφάσεων, όπου κάθε ενέργεια επηρεάζει τις

μελλοντικές καταστάσεις, τις πιθανές ανταμοιβές και τη συνολική στρατηγική του πράκτορα (Sutton & Barto, 2018).

Η εξέλιξη αυτή σηματοδοτεί τη μετάβαση από απλά συστήματα κανόνων σε ευφυή συστήματα αυτόνομης λήψης αποφάσεων, τα οποία μπορούν να προσαρμόζονται δυναμικά στις μεταβαλλόμενες συνθήκες της αγοράς. Στο πλαίσιο αυτό, η σύγχρονη αλγοριθμική διαπραγμάτευση δεν περιορίζεται μόνο στην πρόβλεψη τιμών, αλλά επεκτείνεται στη βελτιστοποίηση σύνθετων στρατηγικών μέσω αξιοποίησης πολυπαραγοντικής πληροφορίας αγοράς, υψηλής χρονικής ανάλυσης δεδομένων και προηγμένων μηχανισμών μάθησης.

2.2 Τεχνική ανάλυση και βασικοί δείκτες

Η τεχνική ανάλυση αποτέλεσε μία από τις πρώτες συστηματικές προσεγγίσεις στη μελέτη των χρηματοοικονομικών αγορών και εξακολουθεί να χρησιμοποιείται ευρέως τόσο από ιδιώτες όσο και από θεσμικούς επενδυτές. Βασίζεται στην υπόθεση ότι η ιστορική συμπεριφορά των τιμών και του όγκου συναλλαγών ενσωματώνει σημαντικό μέρος της διαθέσιμης πληροφορίας της αγοράς και ότι επαναλαμβανόμενα μοτίβα μπορούν να αξιοποιηθούν για την εκτίμηση πιθανών μελλοντικών κινήσεων (Murphy, 1999).

Στο πλαίσιο αυτό, έχουν αναπτυχθεί πολυάριθμοι τεχνικοί δείκτες με στόχο τη μέτρηση της τάσης, της ορμής, της μεταβλητότητας και της σχετικής ισχύος της αγοράς. Ο Relative Strength Index (RSI) χρησιμοποιείται για την εκτίμηση καταστάσεων υπεραγοράς ή υπερπώλησης, βασιζόμενος στην ένταση και την ταχύτητα των πρόσφατων μεταβολών της τιμής. Ο Moving Average Convergence Divergence (MACD) συγκρίνει δύο εκθετικούς κινητούς μέσους όρους, παρέχοντας ενδείξεις πιθανών αλλαγών στην κατεύθυνση της τάσης. Παράλληλα, δείκτες όπως ο Stochastic Oscillator και ο Schaff Trend Cycle επιχειρούν να αποτυπώσουν την κυκλικότητα της αγοράς και τα πιθανά σημεία αντιστροφής της τάσης (Schaff, 2008).

Οι δείκτες αυτοί αποτέλεσαν τη βάση για την ανάπτυξη πολλών πρώιμων αλγοριθμικών στρατηγικών, καθώς προσφέρονται για σχετικά απλή αυτοματοποίηση και ερμηνεία. Παράλληλα, συνέβαλαν σημαντικά στη διαμόρφωση δομημένων προσεγγίσεων ανάλυσης αγοράς, επιτρέποντας τη συστηματική αξιολόγηση δεδομένων τιμών και όγκου.

Ωστόσο, οι παραδοσιακοί τεχνικοί δείκτες παρουσιάζουν σημαντικούς περιορισμούς. Οι παράμετροί τους είναι συνήθως στατικές και απαιτούν χειροκίνητη ρύθμιση, γεγονός που περιορίζει την προσαρμοστικότητά τους σε μεταβαλλόμενες συνθήκες αγοράς. Επιπλέον, η αποτελεσματικότητά τους μπορεί να μειώνεται σε περιβάλλοντα υψηλής μεταβλητότητας, έντονου θορύβου ή μη γραμμικής συμπεριφοράς. Οι περισσότεροι δείκτες λειτουργούν απομονωμένα, χωρίς ενσωματωμένο μηχανισμό μάθησης, προσαρμογής ή συνδυαστικής αξιοποίησης πολλαπλών πληροφοριακών πηγών.

Ως αποτέλεσμα, αν και η τεχνική ανάλυση προσφέρει σημαντικά πλεονεκτήματα ως προς την ερμηνευσιμότητα και τη σχετική απλότητα εφαρμογής, αδυνατεί συχνά να αντιμετωπίσει πλήρως τη δυναμική και στοχαστική φύση των σύγχρονων χρηματοοικονομικών αγορών. Οι περιορισμοί αυτοί οδήγησαν στην αναζήτηση πιο ευέλικτων και προσαρμοστικών προσεγγίσεων, όπως τα στατιστικά μοντέλα, η μηχανική μάθηση και οι μέθοδοι ενισχυτικής μάθησης, οι οποίες επιτρέπουν την αξιοποίηση πολυπαραγοντικής πληροφορίας και τη δυναμική προσαρμογή της στρατηγικής με βάση τα δεδομένα.

2.3 Στατιστικά μοντέλα χρονοσειρών (ARIMA, GARCH)

Η χρήση στατιστικών μοντέλων χρονοσειρών αποτέλεσε σημαντικό επόμενο βήμα στην ανάλυση χρηματοοικονομικών δεδομένων, προσφέροντας αυστηρό μαθηματικό πλαίσιο για την περιγραφή, μοντελοποίηση και πρόβλεψη των αγορών. Τα μοντέλα AutoRegressive Integrated Moving Average (ARIMA) χρησιμοποιούνται ευρέως για την ανάλυση χρονοσειρών, λαμβάνοντας υπόψη αυτοσυσχετίσεις, τάσεις και πιθανές εποχικές συνιστώσες (Hyndman, 2011). Τα μοντέλα αυτά επιτρέπουν τη βραχυπρόθεσμη πρόβλεψη υπό την παραδοχή σχετικής σταθερότητας των στατιστικών ιδιοτήτων της σειράς.

Παράλληλα, τα μοντέλα Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) αναπτύχθηκαν για την αποτύπωση της χρονικά μεταβαλλόμενης μεταβλητότητας, η οποία αποτελεί θεμελιώδες χαρακτηριστικό των χρηματοοικονομικών αγορών. Τα GARCH μοντέλα είναι ιδιαίτερα χρήσιμα στη διαχείριση κινδύνου, στην εκτίμηση αβεβαιότητας και στην ανάλυση περιόδων έντονης ή περιορισμένης μεταβλητότητας, παρέχοντας πιο ρεαλιστική περιγραφή της συμπεριφοράς των αγορών σε σχέση με απλούστερα γραμμικά μοντέλα.

Παρά τα σημαντικά πλεονεκτήματά τους, τα στατιστικά μοντέλα χρονοσειρών παρουσιάζουν ουσιώδεις περιορισμούς όταν εφαρμόζονται σε σύνθετα και δυναμικά χρηματοοικονομικά περιβάλλοντα. Συχνά βασίζονται σε υποθέσεις γραμμικότητας, στασιμότητας ή σχετικά σταθερών σχέσεων, οι οποίες ενδέχεται να παραβιάζονται σε πραγματικές αγορές, ιδιαίτερα σε ενδοημερήσια δεδομένα υψηλής συχνότητας. Επιπλέον, τα μοντέλα αυτά επικεντρώνονται κυρίως στην πρόβλεψη τιμών ή μεταβλητότητας και όχι στη βελτιστοποίηση πραγματικών στρατηγικών διαδοχικής λήψης αποφάσεων.

Ως αποτέλεσμα, αν και τα μοντέλα ARIMA και GARCH προσφέρουν σημαντική θεωρητική συνέπεια και πολύτιμα εργαλεία ποσοτικής ανάλυσης, δεν επαρκούν από μόνα τους για την ανάπτυξη πλήρως προσαρμοστικών συστημάτων αυτόματης διαπραγμάτευσης. Οι περιορισμοί αυτοί ενίσχυσαν τη μετάβαση προς πιο ευέλικτες προσεγγίσεις, όπως η μηχανική μάθηση και η ενισχυτική μάθηση, οι οποίες αντιμετωπίζουν τη χρηματοοικονομική διαπραγμάτευση όχι μόνο ως πρόβλημα πρόβλεψης, αλλά ως διαδικασία στρατηγικής αλληλεπίδρασης με ένα δυναμικό περιβάλλον αγοράς.

2.4 Εφαρμογές μηχανικής μάθησης στις χρηματοοικονομικές αγορές

Η μηχανική μάθηση αποτέλεσε καθοριστικό βήμα στην εξέλιξη των μεθόδων ανάλυσης και πρόβλεψης των χρηματοοικονομικών αγορών, καθώς επέτρεψε την υπέρβαση πολλών περιορισμών των παραδοσιακών στατιστικών μοντέλων. Σε αντίθεση με τις κλασικές προσεγγίσεις, τα μοντέλα μηχανικής μάθησης δεν απαιτούν αυστηρές παραδοχές γραμμικότητας ή στασιμότητας και μπορούν να μάθουν απευθείας από τα δεδομένα, αναγνωρίζοντας πολύπλοκες και μη γραμμικές σχέσεις.

Στη σχετική βιβλιογραφία έχουν εφαρμοστεί ποικίλες τεχνικές μηχανικής μάθησης για χρηματοοικονομική πρόβλεψη, όπως πολυεπίπεδα νευρωνικά δίκτυα, Support Vector Machines και σύγχρονες αρχιτεκτονικές βαθιάς μάθησης. Ιδιαίτερη σημασία παρουσιάζουν τα επαναληπτικά νευρωνικά δίκτυα και ειδικότερα τα Long Short-Term Memory (LSTM), τα οποία είναι σχεδιασμένα για την επεξεργασία ακολουθιακών δεδομένων και μπορούν να αποτυπώνουν μακροχρόνιες χρονικές εξαρτήσεις στις χρονοσειρές τιμών (Li et al., 2023; Van Houdt et al., 2020).

Πέρα από τη βασική αρχιτεκτονική, η αποτελεσματικότητα των μοντέλων μηχανικής μάθησης επηρεάζεται σημαντικά από την επιλογή υπερπαραμέτρων, συναρτήσεων ενεργοποίησης και διαδικασιών βελτιστοποίησης. Οι Sami et al. (2023) έδειξαν ότι η επιλογή διαφορετικών συναρτήσεων ενεργοποίησης μπορεί να επηρεάσει ουσιαστικά την ακρίβεια πρόβλεψης σε χρηματοοικονομικά δεδομένα, αναδεικνύοντας τη σημασία της λεπτομερούς παραμετροποίησης πέρα από τη βασική δομή του μοντέλου.

Μελέτες όπως των Kenniy et al. (2022) και Jinshimo et al. (2023) επιβεβαιώνουν ότι τα μοντέλα βαθιάς μάθησης συχνά επιτυγχάνουν βελτιωμένη απόδοση πρόβλεψης σε σχέση με παραδοσιακά στατιστικά μοντέλα, ιδιαίτερα όταν εφαρμόζονται σε δεδομένα υψηλής διάστασης και έντονων χρονικών εξαρτήσεων. Ωστόσο, η αυξημένη πολυπλοκότητα αυτών των μοντέλων συνοδεύεται από προκλήσεις όπως ο κίνδυνος υπερεκπαίδευσης, η ανάγκη εκτεταμένων υπολογιστικών πόρων και η περιορισμένη ερμηνευσιμότητα.

Ένα βασικό μειονέκτημα των περισσότερων εφαρμογών μηχανικής και βαθιάς μάθησης στις χρηματοοικονομικές αγορές είναι ότι επικεντρώνονται κυρίως στο πρόβλημα της πρόβλεψης της μελλοντικής τιμής ή κατεύθυνσης της αγοράς. Παρότι η πρόβλεψη αυτή αποτελεί χρήσιμο εργαλείο, δεν μεταφράζεται άμεσα σε ολοκληρωμένες στρατηγικές διαπραγμάτευσης, καθώς δεν ενσωματώνει ρητά τη διαδικασία διαδοχικής λήψης αποφάσεων, τη δυναμική εξέλιξη του χαρτοφυλακίου ή τη στρατηγική προσαρμογή σε πραγματικό χρόνο.

Οι περιορισμοί αυτοί ανέδειξαν την ανάγκη για μεθόδους που δεν περιορίζονται αποκλειστικά στην πρόβλεψη, αλλά ενσωματώνουν τη διαδικασία λήψης αποφάσεων, στρατηγικής προσαρμογής και βελτιστοποίησης σε ένα ενιαίο πλαίσιο. Η ανάγκη αυτή οδήγησε στη μετάβαση προς τις μεθόδους ενισχυτικής μάθησης, οι οποίες αντιμετωπίζουν τη χρηματοοικονομική διαπραγμάτευση ως πρόβλημα στρατηγικής αλληλεπίδρασης με το περιβάλλον αγοράς.

2.5 Εισαγωγή στην ενισχυτική μάθηση (Reinforcement Learning)

Η ενισχυτική μάθηση (Reinforcement Learning – RL) αποτελεί διακριτό παράδειγμα μηχανικής μάθησης, το οποίο επικεντρώνεται στη βελτιστοποίηση διαδοχικών αποφάσεων μέσω αλληλεπίδρασης με ένα δυναμικό περιβάλλον. Σε αντίθεση με την επιβλεπόμενη

μάθηση, όπου το μοντέλο εκπαιδεύεται με γνωστά ζεύγη εισόδου–εξόδου, και τη μη επιβλεπόμενη μάθηση, όπου αναζητούνται εσωτερικές δομές στα δεδομένα, η ενισχυτική μάθηση βασίζεται στη σταδιακή απόκτηση εμπειρίας και στην ανατροφοδότηση μέσω ανταμοιβών (Sutton & Barto, 2018).

Το βασικό πλαίσιο της ενισχυτικής μάθησης περιλαμβάνει έναν πράκτορα (agent), ο οποίος αλληλεπιδρά με ένα περιβάλλον (environment). Σε κάθε χρονικό βήμα, ο πράκτορας παρατηρεί την τρέχουσα κατάσταση του περιβάλλοντος, επιλέγει μια ενέργεια από ένα προκαθορισμένο σύνολο ενεργειών και λαμβάνει ανταμοιβή, ενώ το περιβάλλον μεταβαίνει σε νέα κατάσταση. Στόχος του πράκτορα είναι η εκμάθηση μιας πολιτικής (policy), δηλαδή ενός κανόνα επιλογής ενεργειών, που μεγιστοποιεί το αναμενόμενο άθροισμα μελλοντικών ανταμοιβών. Η επιλογή της συνάρτησης ανταμοιβής είναι καθοριστικής σημασίας, καθώς επηρεάζει άμεσα τη συμπεριφορά και τη στρατηγική που θα αναπτύξει ο πράκτορας.

Η θεωρητική βάση της ενισχυτικής μάθησης στηρίζεται στη μοντελοποίηση του προβλήματος ως Markov Decision Process (MDP), όπου η μελλοντική εξέλιξη εξαρτάται από την τρέχουσα κατάσταση και την επιλεγμένη ενέργεια. Η προσέγγιση αυτή είναι ιδιαίτερα κατάλληλη για χρηματοοικονομικές εφαρμογές, καθώς επιτρέπει την αναπαράσταση της αγοράς και της κατάστασης του χαρτοφυλακίου ως διαδοχικών εξελισσόμενων καταστάσεων. Σε τέτοια περιβάλλοντα, η κατάσταση του συστήματος αναπαρίσταται συνήθως μέσω πολυδιάστατων διανυσμάτων χαρακτηριστικών, τα οποία περιλαμβάνουν πληροφορία από τιμές, όγκο συναλλαγών, τεχνικούς δείκτες και στοιχεία σχετιζόμενα με την υφιστάμενη θέση του πράκτορα.

Σε αντίθεση με πολλά παραδοσιακά μοντέλα μηχανικής μάθησης που επικεντρώνονται κυρίως στην πρόβλεψη τιμών, η ενισχυτική μάθηση αντιμετωπίζει τη διαπραγμάτευση ως πρόβλημα στρατηγικής λήψης αποφάσεων. Ο πράκτορας καλείται να επιλέξει πότε να αγοράσει, να πουλήσει ή να διακρατήσει μια θέση, λαμβάνοντας υπόψη όχι μόνο την τρέχουσα πληροφορία αγοράς αλλά και τις μακροπρόθεσμες συνέπειες των ενεργειών του στη συνολική απόδοση του χαρτοφυλακίου.

Η ενισχυτική μάθηση έχει εφαρμοστεί σε πληθώρα χρηματοοικονομικών προβλημάτων, όπως αυτόματη διαπραγμάτευση, διαχείριση χαρτοφυλακίου και βελτιστοποίηση στρατηγικών συναλλαγών. Οι Moody και Saffell (1998) υπήρξαν από τους πρώτους που ανέδειξαν τη σημασία συναρτήσεων απόδοσης σε trading συστήματα βασισμένα σε RL, ενώ νεότερες

μελέτες δείχνουν ότι οι πράκτορες ενισχυτικής μάθησης μπορούν να προσαρμόζονται αποτελεσματικά σε μεταβαλλόμενες συνθήκες αγοράς και να επιτυγχάνουν ανταγωνιστική ή ανώτερη απόδοση έναντι στατικών στρατηγικών (Théate & Ernst, 2021; Jinshimo et al., 2023).

Παρά τα πλεονεκτήματά της, η ενισχυτική μάθηση παρουσιάζει και σημαντικές προκλήσεις. Η εκπαίδευση πρακτόρων απαιτεί μεγάλο όγκο δεδομένων, η διαδικασία μάθησης μπορεί να είναι ασταθής και η σωστή σχεδίαση της συνάρτησης ανταμοιβής είναι κρίσιμη. Στα χρηματοοικονομικά περιβάλλοντα, τα οποία χαρακτηρίζονται από θόρυβο, μη στασιμότητα και υψηλή μεταβλητότητα, οι δυσκολίες αυτές εντείνονται.

Οι προκλήσεις αυτές οδήγησαν στην ανάπτυξη μεθόδων βαθιάς ενισχυτικής μάθησης (Deep Reinforcement Learning), όπου οι συναρτήσεις αξίας ή πολιτικής προσεγγίζονται μέσω βαθιών νευρωνικών δικτύων. Η σύζευξη βαθιάς μάθησης και ενισχυτικής μάθησης επέτρεψε την εφαρμογή RL μεθόδων σε περιβάλλοντα υψηλής πολυπλοκότητας και μεγάλου χώρου καταστάσεων, δημιουργώντας τη βάση για προηγμένους αλγορίθμους όπως τα Deep Q-Networks, τα οποία αποτελούν θεμελιώδη άξονα της παρούσας εργασίας.

2.6 Αρχές και αλγόριθμοι Deep Q-Networks

Η ανάπτυξη των Deep Q-Networks (DQN) αποτέλεσε καθοριστικό σημείο στην εξέλιξη της ενισχυτικής μάθησης, καθώς επέτρεψε την εφαρμογή της σε προβλήματα με μεγάλο και πολύπλοκο χώρο καταστάσεων. Στα κλασικά σχήματα Q-learning, η συνάρτηση αξίας δράσης Q(s,a) εκφράζει την αναμενόμενη συνολική ανταμοιβή όταν ο πράκτορας βρίσκεται σε κατάσταση s και επιλέγει την ενέργεια a. Στο πλαίσιο αυτό, το s αντιστοιχεί στην κατάσταση του περιβάλλοντος (state), ενώ το a αντιστοιχεί στην επιλεγμένη ενέργεια (action). Η προσέγγιση αυτή είναι εφαρμόσιμη μόνο σε απλά προβλήματα μικρής κλίμακας και καθίσταται πρακτικά ανεφάρμοστη σε περιβάλλοντα υψηλής διάστασης, όπως οι χρηματοοικονομικές αγορές.

Για την αντιμετώπιση αυτού του περιορισμού, οι Mnih et al. (2015) εισήγαγαν τα Deep Q-Networks, στα οποία η συνάρτηση Q(s,a) προσεγγίζεται από ένα βαθύ νευρωνικό δίκτυο με παραμέτρους θ, όπου θ συμβολίζει το σύνολο των παραμέτρων του δικτύου, δηλαδή τα βάρη και τις μεταβλητές που προσαρμόζονται κατά τη διαδικασία μάθησης, ώστε να

βελτιστοποιείται η προσέγγιση της συνάρτησης αξίας δράσης. Το δίκτυο λαμβάνει ως είσοδο την τρέχουσα κατάσταση του περιβάλλοντος και παράγει ως έξοδο εκτιμήσεις της αξίας Q για κάθε διαθέσιμη ενέργεια. Με τον τρόπο αυτό, η εκμάθηση της βέλτιστης πολιτικής μετατρέπεται σε πρόβλημα βελτιστοποίησης των παραμέτρων του νευρωνικού δικτύου.

Η θεωρητική βάση του DQN στηρίζεται στην εξίσωση Bellman:

$$Q^*(s,a) = E[r + \gamma max_{a'} Q^*(s',a')] \quad (2.1)$$

όπου:

Q*(s,a) είναι η βέλτιστη συνάρτηση αξίας δράσης,

s είναι η τρέχουσα κατάσταση (state) του περιβάλλοντος,

a είναι η επιλεγμένη ενέργεια (action),

r είναι η άμεση ανταμοιβή (reward),

γ είναι ο συντελεστής προεξόφλησης (discount factor),

s′ είναι η επόμενη κατάσταση,

a′ είναι μία πιθανή ενέργεια στην επόμενη κατάσταση s′

maxα′Q*(s′,a′) είναι η μέγιστη εκτιμώμενη αξία της επόμενης κατάστασης μεταξύ όλων των δυνατών ενεργειών.

E[·] συμβολίζει την αναμενόμενη τιμή (expected value) ως προς την πιθανή κατανομή των επόμενων καταστάσεων.

Η σχέση αυτή εκφράζει την αρχή της βέλτιστης υποδομής, σύμφωνα με την οποία η συνολική αξία μιας κατάστασης εξαρτάται τόσο από την άμεση ανταμοιβή όσο και από τη βέλτιστη αναμενόμενη μελλοντική απόδοση.

Στην πράξη, το νευρωνικό δίκτυο εκπαιδεύεται ώστε να ελαχιστοποιεί τη διαφορά μεταξύ της τρέχουσας εκτίμησης Q(s,a;θ) και της τιμής-στόχου που προκύπτει από την εξίσωση Bellman. Η διαδικασία αυτή πραγματοποιείται μέσω μεθόδων βελτιστοποίησης βασισμένων σε gradient descent και οδηγεί στη σταδιακή προσέγγιση της βέλτιστης πολιτικής.

Ωστόσο, η άμεση εφαρμογή νευρωνικών δικτύων στο Q-learning παρουσίασε σημαντικά προβλήματα αστάθειας και πιθανής απόκλισης. Τα προβλήματα αυτά σχετίζονται κυρίως με τη χρονική συσχέτιση διαδοχικών εμπειριών, τη συνεχή μεταβολή των στόχων εκπαίδευσης και τη μεγάλη ευαισθησία σε θορυβώδη δεδομένα, δηλαδή δεδομένα που περιλαμβάνουν τυχαίες ή βραχυπρόθεσμες διακυμάνσεις χωρίς ουσιαστική πληροφοριακή αξία, οι οποίες δυσχεραίνουν τη διάκριση πραγματικών προτύπων από παροδικές μεταβολές.

Για την αντιμετώπιση αυτών των ζητημάτων, οι Mnih et al. (2015) εισήγαγαν δύο βασικούς μηχανισμούς σταθεροποίησης. Οι βασικοί αυτοί μηχανισμοί είναι οι εξής:

1. Experience Replay. Οι εμπειρίες του πράκτορα αποθηκεύονται σε μνήμη επανάληψης εμπειριών και δειγματοληπτούνται τυχαία κατά την εκπαίδευση. Η διαδικασία αυτή μειώνει τη χρονική συσχέτιση των δεδομένων, βελτιώνει τη στατιστική αποδοτικότητα και επιτρέπει επαναχρησιμοποίηση προηγούμενων εμπειριών.
2. Target Network. Χρησιμοποιείται ξεχωριστό δίκτυο στόχος για τον υπολογισμό των τιμών-στόχων, το οποίο ενημερώνεται περιοδικά με τις παραμέτρους του κύριου δικτύου. Με τον τρόπο αυτό, σταθεροποιείται η μαθησιακή διαδικασία και περιορίζεται ο κίνδυνος απότομων αποκλίσεων.

Στο πλαίσιο των χρηματοοικονομικών αγορών, τα DQN προσφέρουν σημαντικά πλεονεκτήματα, καθώς επιτρέπουν την ταυτόχρονη αξιοποίηση πολυδιάστατης πληροφορίας αγοράς. Η δυνατότητα αυτή επιτρέπει τη διαμόρφωση στρατηγικών που υπερβαίνουν την απλή πρόβλεψη τιμών και εστιάζουν στη βελτιστοποίηση συνολικών πολιτικών λήψης αποφάσεων.

Παρά τα σημαντικά πλεονεκτήματά τους, τα DQN παρουσιάζουν και ουσιώδεις περιορισμούς. Ένα από τα βασικότερα προβλήματα είναι η υπερεκτίμηση των τιμών Q, η οποία προκύπτει από τη χρήση του ίδιου δικτύου τόσο για την επιλογή όσο και για την αξιολόγηση της βέλτιστης ενέργειας. Το φαινόμενο αυτό μπορεί να οδηγήσει σε υπεραισιόδοξες εκτιμήσεις, μη ρεαλιστικές πολιτικές και ασταθή συμπεριφορά, ιδιαίτερα σε στοχαστικά περιβάλλοντα όπως οι χρηματοοικονομικές αγορές (Hasselt, 2010).

Οι περιορισμοί αυτοί οδήγησαν στην ανάπτυξη βελτιωμένων παραλλαγών, όπως το Double DQN και η αρχιτεκτονική Dueling DQN, οι οποίες στοχεύουν στη βελτίωση της σταθερότητας, της ακρίβειας και της αποδοτικότητας της μάθησης. Η αρχιτεκτονική Dueling

DQN, η οποία αποτελεί βασικό άξονα της παρούσας εργασίας, παρουσιάζεται αναλυτικά στην επόμενη υποενότητα.

2.7 Η αρχιτεκτονική Dueling Deep Q-Network (Dueling DQN)

Η αρχιτεκτονική Dueling Deep Q-Network (Dueling DQN) προτάθηκε από τους Wang et al. (2016) ως σημαντική επέκταση των κλασικών Deep Q-Networks, με στόχο τη βελτίωση της σταθερότητας, της αποδοτικότητας και της ποιότητας μάθησης σε περιβάλλοντα όπου πολλές διαθέσιμες ενέργειες οδηγούν σε παρόμοια αποτελέσματα. Το βασικό θεωρητικό κίνητρο πίσω από την αρχιτεκτονική αυτή είναι ότι σε πολλές καταστάσεις η συνολική αξιολόγηση της κατάστασης είναι σημαντικότερη από τη λεπτομερή διαφοροποίηση μεταξύ παρόμοιων ενεργειών.

Στα κλασικά DQN, η συνάρτηση αξίας δράσης Q(s,a) προσεγγίζεται άμεσα μέσω ενός ενιαίου νευρωνικού δικτύου. Η προσέγγιση αυτή μπορεί να είναι λιγότερο αποδοτική όταν διαφορετικές ενέργειες (όπως Buy, Sell ή Hold) έχουν συγκρίσιμες επιπτώσεις. Σε τέτοιες περιπτώσεις, το μοντέλο δυσκολεύεται να διακρίνει με ακρίβεια μικρές διαφορές, γεγονός που μπορεί να επιβραδύνει τη σύγκλιση και να αυξήσει την αστάθεια.

Η αρχιτεκτονική Dueling DQN αντιμετωπίζει το πρόβλημα αυτό αποσυνθέτοντας τη συνάρτηση Q(s,a) σε δύο διακριτές συνιστώσες:

- τη συνάρτηση αξίας κατάστασης V(s), η οποία εκφράζει τη συνολική ποιότητα της κατάστασης ανεξάρτητα από την επιλεγμένη ενέργεια,
- τη συνάρτηση πλεονεκτήματος A(s,a), η οποία εκφράζει το σχετικό όφελος της επιλογής μιας συγκεκριμένης ενέργειας σε σχέση με τις υπόλοιπες.

Ο επανασυνδυασμός των δύο συναρτήσεων πραγματοποιείται ως εξής:

$$Q(s,a) = V(s) + \left(A(s,a) - \frac{1}{|A|}\sum_{a'} A(s,a')\right) \quad (2.2)$$

όπου:

Q(s,a) είναι η συνολική αξία δράσης στην κατάσταση s για την ενέργεια a,

V(s) είναι η συνολική αξία της κατάστασης s ανεξάρτητα από την ενέργεια,

A(s,a) είναι το πλεονέκτημα της συγκεκριμένης ενέργειας a στην κατάσταση s,

a′ αντιστοιχεί σε κάθε δυνατή ενέργεια στο σύνολο ενεργειών,

A(s,a′) είναι το πλεονέκτημα κάθε πιθανής ενέργειας στην κατάσταση s,

|A| είναι το πλήθος των διαθέσιμων ενεργειών,

$\sum_{a'}A(s,a')$ είναι το άθροισμα των πλεονεκτημάτων όλων των πιθανών ενεργειών,

$(1/|A|)\sum_{a'}A(s,a')$ είναι ο μέσος όρος του πλεονεκτήματος όλων των ενεργειών.

Η αφαίρεση του μέσου όρου του πλεονεκτήματος αποτρέπει αυθαίρετες μετατοπίσεις μεταξύ των V και A, επιτρέποντας στο δίκτυο να μαθαίνει πιο αποτελεσματικά τόσο τη συνολική αξία μιας κατάστασης όσο και τη σχετική σημασία κάθε ενέργειας. Η χρησιμότητα της αρχιτεκτονικής Dueling DQN είναι ιδιαίτερα σημαντική σε χρηματοοικονομικά περιβάλλοντα, όπου συχνά υπάρχουν εκτεταμένες χρονικές περίοδοι στις οποίες πολλές ενέργειες παρουσιάζουν παρόμοιο βραχυπρόθεσμο αποτέλεσμα. Για παράδειγμα, σε περιόδους χαμηλής μεταβλητότητας ή ουδέτερης αγοράς, η σωστή αξιολόγηση της συνολικής κατάστασης μπορεί να είναι σημαντικότερη από τη διαφορά μεταξύ αγοράς, πώλησης ή διακράτησης.

Σύμφωνα με τους Wang et al. (2016), η αρχιτεκτονική Dueling DQN οδηγεί σε ταχύτερη σύγκλιση, βελτιωμένη σταθερότητα, αποτελεσματικότερη μάθηση σε περιβάλλοντα υψηλής διάστασης και μειωμένη ευαισθησία σε θορυβώδεις ανταμοιβές, δηλαδή ανταμοιβές που επηρεάζονται από υψηλή βραχυπρόθεσμη μεταβλητότητα και τυχαίες διακυμάνσεις της αγοράς.

Τα πλεονεκτήματα αυτά επιβεβαιώνονται και από μεταγενέστερες αναλύσεις, όπως του Sewak (2019), οι οποίες αναδεικνύουν την αυξημένη αποδοτικότητα των Dueling αρχιτεκτονικών σε πολύπλοκα προβλήματα ενισχυτικής μάθησης.

Επιπλέον, η αρχιτεκτονική Dueling DQN λειτουργεί συμπληρωματικά με άλλες βελτιώσεις, όπως το Double DQN (Hasselt, 2010), το οποίο περιορίζει το φαινόμενο υπερεκτίμησης των Q-values. Ο συνδυασμός Double και Dueling DQN ενισχύει σημαντικά τη σταθερότητα της μάθησης σε στοχαστικά περιβάλλοντα, όπως οι χρηματοοικονομικές αγορές.

Στη σχετική χρηματοοικονομική βιβλιογραφία, οι αρχιτεκτονικές Dueling DQN αξιοποιούνται κυρίως σε προβλήματα αυτόματης διαπραγμάτευσης, διαχείρισης χαρτοφυλακίου και στρατηγικής λήψης αποφάσεων σε περιβάλλοντα υψηλής μεταβλητότητας. Μελέτες όπως των Théate και Ernst (2021) και Tsantekidis et al. (2021) υποδεικνύουν ότι οι παραλλαγές αυτές παρουσιάζουν αυξημένη προσαρμοστικότητα σε δυναμικά καθεστώτα αγοράς.

Στο πλαίσιο της παρούσας εργασίας, η επιλογή της αρχιτεκτονικής Dueling DQN θεωρείται ιδιαίτερα κατάλληλη, καθώς το πρόβλημα της ενδοημερήσιας διαπραγμάτευσης χαρακτηρίζεται από πολυδιάστατο χώρο καταστάσεων και περιόδους όπου οι διαθέσιμες ενέργειες δεν παρουσιάζουν έντονες διαφοροποιήσεις ως προς το άμεσο αποτέλεσμα. Ο διαχωρισμός της συνολικής αξίας της κατάστασης από το πλεονέκτημα των επιμέρους ενεργειών επιτρέπει στον πράκτορα να αναπτύσσει πιο σταθερές, αποδοτικές και αξιόπιστες στρατηγικές λήψης αποφάσεων.

2.8 Ο δείκτης Volume-Weighted Average Price (VWAP) και οι χρηματοοικονομικές του εφαρμογές

Ο δείκτης Volume-Weighted Average Price (VWAP) αποτελεί ευρέως χρησιμοποιούμενο τεχνικό δείκτη στις σύγχρονες χρηματοοικονομικές αγορές, ο οποίος συνδυάζει πληροφορία τιμής και όγκου συναλλαγών, παρέχοντας μια πιο ολοκληρωμένη εικόνα της συμπεριφοράς

της αγοράς σε συγκεκριμένη χρονική περίοδο. Ο VWAP εκφράζει τη μέση τιμή διαπραγμάτευσης ενός χρηματοοικονομικού τίτλου, σταθμισμένη ως προς τον όγκο συναλλαγών, αποδίδοντας μεγαλύτερη βαρύτητα σε τιμές όπου πραγματοποιείται υψηλότερη συναλλακτική δραστηριότητα.

Μαθηματικά, ο VWAP υπολογίζεται ως:

$$VWAP = \frac{\sum_{i=1}^{n} P_i V_i}{\sum_{i=1}^{n} V_i} \qquad (2.3)$$

όπου:

P_i είναι η τιμή του τίτλου κατά τη χρονική στιγμή ή συναλλαγή i,

V_i είναι ο αντίστοιχος όγκος συναλλαγών,

n είναι ο συνολικός αριθμός χρονικών βημάτων ή συναλλαγών της εξεταζόμενης περιόδου.

$\sum_{i=1}^{n} P_i V_i$ είναι το συνολικό άθροισμα των γινομένων τιμής και όγκου για όλες τις συναλλαγές της περιόδου,

$\sum_{i=1}^{n} V_i$ είναι το συνολικό άθροισμα του όγκου συναλλαγών για την ίδια περίοδο.

Η μαθηματική αυτή σχέση επιτρέπει στον VWAP να αποτυπώνει όχι μόνο τη μέση πορεία της τιμής, αλλά και τη σχετική σημασία των επιμέρους συναλλαγών βάσει της συναλλακτικής έντασης.

Στη χρηματοοικονομική πρακτική, ο VWAP χρησιμοποιείται συχνά ως εργαλείο ανάλυσης της σχέσης τιμής–όγκου, ως σημείο αναφοράς για τη συνολική δυναμική της αγοράς και ως benchmark σε στρατηγικές εκτέλεσης εντολών μεγάλης κλίμακας. Στη σχετική βιβλιογραφία, ο VWAP συνδέεται ιδιαίτερα με προβλήματα βέλτιστης εκτέλεσης εντολών (optimal execution), όπου στόχος είναι η σταδιακή εκτέλεση μεγάλων εντολών με τρόπο που να προσεγγίζει τη μέση σταθμισμένη τιμή της αγοράς (Guéant et al., 2014; Frei et al., 2013).

Οι Humphery-Jenner et al. (2011) και McCulloch et al. (2012) εξετάζουν στρατηγικές βελτιστοποίησης βασισμένες στον VWAP υπό συνθήκες αβεβαιότητας, ενώ πιο πρόσφατες μελέτες, όπως των Xiaodong et al. (2022) και Soohan et al. (2024), επιχειρούν να συνδυάσουν

τον VWAP με τεχνικές ενισχυτικής μάθησης για δυναμική προσαρμογή στρατηγικών εκτέλεσης.

Ωστόσο, στη μεγάλη πλειονότητα των σχετικών εφαρμογών, ο VWAP χρησιμοποιείται κυρίως ως εξωτερικό benchmark αξιολόγησης ή ως εργαλείο εκτέλεσης, και όχι ως άμεσο χαρακτηριστικό εισόδου στη διαδικασία λήψης αποφάσεων ενός πράκτορα μάθησης.

Η παρούσα εργασία διαφοροποιείται ουσιαστικά από αυτή την προσέγγιση. Ο VWAP δεν αντιμετωπίζεται ως κύριο κριτήριο ποιότητας εκτέλεσης, αλλά ως πρόσθετο πληροφοριακό χαρακτηριστικό αγοράς, το οποίο ενσωματώνεται στο διάνυσμα κατάστασης (state representation) του πράκτορα ενισχυτικής μάθησης. Με τον τρόπο αυτό, ο πράκτορας αποκτά άμεση πρόσβαση σε δεδομένα που συνδυάζουν τιμή και όγκο συναλλαγών, επιτρέποντας τη διερεύνηση του κατά πόσο η συμπληρωματική αυτή πληροφορία μπορεί να βελτιώσει τη μαθησιακή διαδικασία και τη συνολική στρατηγική λήψης αποφάσεων.

Η προσέγγιση αυτή μετατοπίζει τον ρόλο του VWAP από παθητικό εργαλείο εκ των υστέρων αξιολόγησης σε ενεργό στοιχείο εμπλουτισμού της αναπαράστασης της αγοράς. Ως αποτέλεσμα, η χρήση του VWAP στην παρούσα εργασία εντάσσεται περισσότερο στο πλαίσιο του feature engineering και της πολυπαραγοντικής αναπαράστασης καταστάσεων παρά στο κλασικό πλαίσιο βελτιστοποίησης εκτέλεσης.

Συνολικά, ο δείκτης VWAP αποτελεί σημαντικό εργαλείο χρηματοοικονομικής ανάλυσης, καθώς συνδυάζει δύο κρίσιμες διαστάσεις της αγοράς — την τιμή και τον όγκο — προσφέροντας πολύτιμη συμπληρωματική πληροφορία για τη δομή και τη δυναμική των συναλλαγών. Η ενσωμάτωσή του σε αρχιτεκτονικές βαθιάς ενισχυτικής μάθησης, όπως η Dueling DQN, δημιουργεί ένα συνεκτικό θεωρητικό πλαίσιο για τη διερεύνηση πιο σύνθετων και προσαρμοστικών στρατηγικών αυτόματης διαπραγμάτευσης.

Συνοψίζοντας, η θεωρητική ανάλυση του Κεφαλαίου 2 αναδεικνύει τα ακόλουθα:

- Η αλγοριθμική διαπραγμάτευση εξελίχθηκε από στατικές και εμπειρικές στρατηγικές σε σύνθετα, αυτοματοποιημένα και προσαρμοστικά συστήματα λήψης αποφάσεων.
- Οι τεχνικοί δείκτες και οι παραδοσιακές μέθοδοι τεχνικής ανάλυσης παρέχουν χρήσιμη πληροφορία αγοράς, αλλά περιορίζονται από στατικότητα και περιορισμένη προσαρμοστικότητα.

- Τα στατιστικά μοντέλα χρονοσειρών προσφέρουν θεωρητική συνέπεια, αλλά δυσκολεύονται να ανταποκριθούν στη μη γραμμικότητα και τη μη στασιμότητα των σύγχρονων αγορών.
- Οι μέθοδοι μηχανικής και βαθιάς μάθησης ενισχύουν σημαντικά την ικανότητα πρόβλεψης, χωρίς όμως να επιλύουν πλήρως το πρόβλημα της στρατηγικής διαδοχικής λήψης αποφάσεων.
- Η ενισχυτική μάθηση εισάγει ένα ολοκληρωμένο πλαίσιο στρατηγικής βελτιστοποίησης, επιτρέποντας στον πράκτορα να μαθαίνει πολιτικές που μεγιστοποιούν μακροπρόθεσμη απόδοση μέσω αλληλεπίδρασης με το περιβάλλον.
- Οι αρχιτεκτονικές Deep Q-Networks και ειδικότερα η Dueling DQN επιτρέπουν την αποτελεσματική εφαρμογή βαθιάς ενισχυτικής μάθησης σε περιβάλλοντα υψηλής πολυπλοκότητας, βελτιώνοντας τη σταθερότητα και την αποδοτικότητα της μάθησης.
- Ο δείκτης VWAP αποτελεί σημαντικό πληροφοριακό εργαλείο που συνδυάζει τιμή και όγκο συναλλαγών. Στην παρούσα εργασία αξιοποιείται ως πρόσθετο χαρακτηριστικό εισόδου στο διάνυσμα κατάστασης του πράκτορα, επιτρέποντας τη διερεύνηση της συμβολής του στον εμπλουτισμό της διαδικασίας λήψης αποφάσεων.

3. Μεθοδολογία

3.1 Ερευνητική προσέγγιση και γενικός σχεδιασμός

Η παρούσα εργασία ακολουθεί ερευνητική προσέγγιση βασισμένη στη σχεδίαση, ανάπτυξη και αξιολόγηση ενός υπολογιστικού συστήματος αυτόματης ενδοημερήσιας διαπραγμάτευσης, το οποίο αξιοποιεί τεχνικές βαθιάς ενισχυτικής μάθησης για τη λήψη διαδοχικών επενδυτικών αποφάσεων. Η επιλογή της συγκεκριμένης προσέγγισης προκύπτει από τη θεωρητική ανάλυση του προηγούμενου κεφαλαίου, όπου αναδείχθηκε ότι οι παραδοσιακές προβλεπτικές και στατικές μέθοδοι παρουσιάζουν σημαντικούς περιορισμούς ως προς την αντιμετώπιση της δυναμικής, πολυπαραγοντικής και στοχαστικής φύσης των χρηματοοικονομικών αγορών.

Η μεθοδολογική στρατηγική δεν περιορίζεται στην πρόβλεψη μελλοντικών τιμών, αλλά αντιμετωπίζει τη διαπραγμάτευση ως πρόβλημα διαδοχικής λήψης αποφάσεων υπό αβεβαιότητα. Σε κάθε χρονικό βήμα, ο πράκτορας καλείται να επιλέξει μεταξύ ενεργειών αγοράς (buy), πώλησης (sell) ή αναμονής (hold), με κάθε επιλογή να επηρεάζει όχι μόνο το άμεσο αποτέλεσμα αλλά και τη μελλοντική κατάσταση του χαρτοφυλακίου και τις επόμενες διαθέσιμες στρατηγικές επιλογές. Για τον λόγο αυτό, η ενισχυτική μάθηση επιλέγεται ως το καταλληλότερο θεωρητικό και πρακτικό πλαίσιο.

Ο γενικός σχεδιασμός της μεθοδολογίας βασίζεται στην ανάπτυξη ενός πράκτορα ενισχυτικής μάθησης, ο οποίος αλληλεπιδρά με ένα προσομοιωμένο περιβάλλον ενδοημερήσιας διαπραγμάτευσης βασισμένο σε ιστορικά δεδομένα αγοράς υψηλής χρονικής ανάλυσης. Το περιβάλλον αναπαριστά τη δυναμική της αγοράς μέσω δεδομένων τιμών και όγκου συναλλαγών σε χρονική ανάλυση επιπέδου λεπτού και παρέχει στον πράκτορα τις απαραίτητες πληροφορίες για τη διαμόρφωση αποφάσεων. Σε κάθε χρονικό βήμα, ο πράκτορας παρατηρεί την κατάσταση της αγοράς, επιλέγει ενέργεια και λαμβάνει ανταμοιβή, η οποία συνδέεται με το οικονομικό αποτέλεσμα των ενεργειών του και διαμορφώνεται με τρόπο που λαμβάνει υπόψη τόσο την απόδοση όσο και στοιχεία διαχείρισης κινδύνου.

Κεντρικό στοιχείο της ερευνητικής προσέγγισης αποτελεί η επιλογή της αρχιτεκτονικής Dueling Deep Q-Network ως βασικού αλγορίθμου μάθησης. Η επιλογή αυτή δικαιολογείται από την ανάγκη σταθερής και αποδοτικής εκμάθησης σε περιβάλλοντα όπου διαφορετικές ενέργειες παρουσιάζουν συχνά παρόμοια βραχυπρόθεσμα αποτελέσματα. Μέσω του

διαχωρισμού της συνολικής αξίας κατάστασης από το πλεονέκτημα των επιμέρους ενεργειών, η αρχιτεκτονική Dueling DQN επιτρέπει την ανάπτυξη πιο σταθερών, γενικεύσιμων και αποδοτικών πολιτικών.

Ιδιαίτερη έμφαση δίνεται επίσης στην αξιοποίηση του δείκτη Volume-Weighted Average Price (VWAP) ως πρόσθετου χαρακτηριστικού εισόδου στο διάνυσμα κατάστασης του πράκτορα. Στην παρούσα εργασία, ο VWAP δεν χρησιμοποιείται ως μηχανισμός βελτιστοποίησης εκτέλεσης, αλλά ως συμπληρωματικό πληροφοριακό χαρακτηριστικό που ενισχύει την αναπαράσταση της αγοράς, επιτρέποντας τη διερεύνηση της συμβολής του στην ποιότητα της μαθησιακής διαδικασίας και στη συνολική απόδοση της στρατηγικής.

Η ερευνητική διαδικασία οργανώνεται σε διακριτά στάδια:

- ορισμός του περιβάλλοντος συναλλαγών, των καταστάσεων, των ενεργειών και της συνάρτησης ανταμοιβής,
- σχεδιασμός και υλοποίηση του πράκτορα Dueling DQN,
- ενσωμάτωση πρόσθετων χαρακτηριστικών εισόδου, συμπεριλαμβανομένου του VWAP,
- εκπαίδευση του πράκτορα σε ιστορικά δεδομένα αγοράς,
- αξιολόγηση μέσω δύο διακριτών πειραματικών σεναρίων διαφορετικής χρονικής κλίμακας,
- συγκριτική αποτίμηση της απόδοσης μεταξύ των εκδοχών BASE και VWAP.

Η αξιολόγηση πραγματοποιείται μέσω δύο βασικών πειραματικών σεναρίων, τα οποία διαφοροποιούνται ως προς τη χρονική κλίμακα, τη διάρθρωση των δεδομένων και τη διαδικασία εκπαίδευσης και ελέγχου, επιτρέποντας τόσο διερευνητική όσο και αυστηρότερη συγκριτική αξιολόγηση της προτεινόμενης προσέγγισης.

Συνολικά, η ερευνητική προσέγγιση της παρούσας εργασίας συνδυάζει θεωρητική τεκμηρίωση, μεθοδολογική αυστηρότητα και πρακτική υλοποίηση, επιδιώκοντας να γεφυρώσει τη θεωρητική έρευνα στη βαθιά ενισχυτική μάθηση με ρεαλιστικές εφαρμογές αυτόματης χρηματοοικονομικής διαπραγμάτευσης. Ο σχεδιασμός αυτός θέτει το θεμέλιο για την αναλυτική παρουσίαση της αρχιτεκτονικής του συστήματος, του περιβάλλοντος εκπαίδευσης και των πειραματικών διαδικασιών που ακολουθούν στις επόμενες ενότητες.

3.2 Αρχιτεκτονική του προτεινόμενου συστήματος

Η αρχιτεκτονική του προτεινόμενου συστήματος βασίζεται στο πρότυπο agent–environment της ενισχυτικής μάθησης και υλοποιεί το πρόβλημα της αλγοριθμικής διαπραγμάτευσης ως διαδικασία διαδοχικών αποφάσεων σε διακριτό χρόνο. Το σύστημα οργανώνεται γύρω από δύο βασικά δομικά στοιχεία, το περιβάλλον διαπραγμάτευσης (trading environment) και τον πράκτορα ενισχυτικής μάθησης (agent), τα οποία αλληλεπιδρούν μέσω της συνάρτησης ανταμοιβής (reward mechanism).

Το περιβάλλον διαπραγμάτευσης προσομοιώνει τη δυναμική της αγοράς σε ενδοημερήσιο επίπεδο και οργανώνεται σε επεισόδια, όπου κάθε επεισόδιο αντιστοιχεί σε μία πλήρη χρηματιστηριακή συνεδρίαση. Σε κάθε χρονικό βήμα t, το περιβάλλον παρέχει στον πράκτορα μια κατάσταση s_t (state), η οποία περιλαμβάνει τόσο πληροφορία αγοράς όσο και πληροφορία σχετική με τη θέση και την προηγούμενη συμπεριφορά του πράκτορα. Γενικά, η κατάσταση μπορεί να αναπαρασταθεί ως:

$$s_t = [\varphi_1(t),\ \varphi_2(t), \dots, \varphi_n(t)] \qquad (3.1)$$

όπου κάθε επιμέρους φi(t) αντιστοιχεί σε πληροφορία σχετική με συγκεκριμένο χρηματοοικονομικό χαρακτηριστικό, όπως αποδόσεις, μεταβολές όγκου συναλλαγών, τεχνικούς δείκτες, χαρακτηριστικά χαρτοφυλακίου και θέσης, καθώς και, όπου εφαρμόζεται, χαρακτηριστικά που βασίζονται στον δείκτη VWAP, συμπεριλαμβανομένης της τιμής VWAP και της σχετικής απόκλισης της τρέχουσας τιμής από αυτόν. Η μαθηματική αυτή αναπαράσταση επιτρέπει τυπικό ορισμό του πληροφοριακού χώρου του πράκτορα και διευκολύνει τη συγκριτική αξιολόγηση διαφορετικών εκδοχών του συστήματος.

Ο πράκτορας επιλέγει σε κάθε χρονικό βήμα μία ενέργεια a_t από διακριτό χώρο ενεργειών:

$$a_t \in \{Hold, Buy, Sell\} \qquad (3.2)$$

όπου η ενέργεια μπορεί να αντιστοιχεί σε διακράτηση, αγορά ή πώληση για το εκάστοτε περιουσιακό στοιχείο. Ο ορισμός διακριτού χώρου ενεργειών καθιστά το πρόβλημα υπολογιστικά διαχειρίσιμο, ενώ επιτρέπει σαφέστερη ερμηνεία της στρατηγικής που μαθαίνει ο πράκτορας.

Η ροή πληροφορίας στο σύστημα ακολουθεί επαναληπτική διαδικασία κλειστού βρόχου (closed-loop). Η κατάσταση του περιβάλλοντος τροφοδοτείται ως είσοδος στο νευρωνικό

δίκτυο του πράκτορα, το οποίο υπολογίζει εκτιμήσεις της συνάρτησης αξίας δράσης Q(s,a) για όλες τις δυνατές ενέργειες. Με βάση τις εκτιμήσεις αυτές και την πολιτική εξερεύνησης, επιλέγεται η ενέργεια που θα εφαρμοστεί στο περιβάλλον. Η ενέργεια αυτή επηρεάζει την κατάσταση του χαρτοφυλακίου και, σε συνδυασμό με την εξέλιξη των τιμών της αγοράς, οδηγεί στον υπολογισμό της ανταμοιβής. Η νέα κατάσταση και η ανταμοιβή χρησιμοποιούνται στη συνέχεια για την ενημέρωση της συνάρτησης αξίας δράσης μέσω της διαδικασίας μάθησης.

Ο πράκτορας υλοποιείται με βάση την αρχιτεκτονική Dueling Deep Q-Network (Dueling DQN), όπου η συνάρτηση αξίας δράσης προσεγγίζεται μέσω ενός βαθύ νευρωνικού δικτύου. Το δίκτυο λαμβάνει ως είσοδο την κατάσταση του περιβάλλοντος και παράγει εκτιμήσεις για την αναμενόμενη μελλοντική ανταμοιβή κάθε δυνατής ενέργειας. Η εκπαίδευση πραγματοποιείται με βάση τον αλγόριθμο Q-learning, τη χρήση μηχανισμού επανάληψης εμπειριών (experience replay), target network και πολιτική εξερεύνησης τύπου ε-greedy, ώστε να επιτυγχάνεται ισορροπία μεταξύ εξερεύνησης και εκμετάλλευσης.

Στην αρχιτεκτονική Dueling DQN, η συνάρτηση αξίας δράσης Q(s,a) αποσυντίθεται σε συνιστώσες αξίας κατάστασης V(s) και πλεονεκτήματος δράσης A(s,a), σύμφωνα με τη θεωρητική διατύπωση που παρουσιάστηκε στο προηγούμενο κεφάλαιο. Ο διαχωρισμός αυτός επιτρέπει πιο αποδοτική εκμάθηση σε περιβάλλοντα όπου οι διαθέσιμες ενέργειες παρουσιάζουν συχνά παρόμοια αποτελέσματα, βελτιώνοντας τη σταθερότητα και τη γενικευσιμότητα του μοντέλου.

Ιδιαίτερη σημασία δίνεται στις σχεδιαστικές επιλογές που αφορούν τον ορισμό της χρονικής δομής του προβλήματος. Η επιλογή κάθε επεισοδίου ως ανεξάρτητης ημερήσιας συνεδρίασης αντανακλά τη λειτουργική πραγματικότητα των ενδοημερήσιων συναλλαγών και επιτρέπει τη συστηματική σύγκριση της απόδοσης του πράκτορα σε ομοιογενείς χρονικές περιόδους.

Η αρχιτεκτονική του συστήματος είναι πλήρως αρθρωτή (modular). Το περιβάλλον διαπραγμάτευσης, ο πράκτορας, η συνάρτηση ανταμοιβής και η διαδικασία προεπεξεργασίας υλοποιούνται ως διακριτές συνιστώσες, επιτρέποντας την ανεξάρτητη τροποποίηση ή αντικατάστασή τους. Η προσέγγιση αυτή διευκολύνει τη διερεύνηση εναλλακτικών σχεδιαστικών επιλογών, όπως διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων ή διαφορετικοί συνδυασμοί χαρακτηριστικών εισόδου, συμπεριλαμβανομένης της

ενσωμάτωσης του δείκτη VWAP στο διάνυσμα κατάστασης, χωρίς να απαιτείται συνολικός ανασχεδιασμός του συστήματος.

Συνολικά, η προτεινόμενη αρχιτεκτονική παρέχει ένα σταθερό, ευέλικτο και επιστημονικά τεκμηριωμένο και μεθοδολογικά συνεπές πλαίσιο για την ανάπτυξη, εκπαίδευση και συγκριτική αξιολόγηση στρατηγικών βαθιάς ενισχυτικής μάθησης σε περιβάλλοντα αυτόματης χρηματοοικονομικής διαπραγμάτευσης.

3.3 Συνάρτηση αξίας δράσης Q(s,a) και συνάρτηση ανταμοιβής

Κεντρικό στοιχείο της μεθοδολογίας της παρούσας εργασίας αποτελεί ο ορισμός της συνάρτησης αξίας δράσης Q(s,a) και της συνάρτησης ανταμοιβής (reward function), καθώς οι δύο αυτές συνιστώσες καθορίζουν άμεσα τη μαθησιακή συμπεριφορά, τη στρατηγική προσαρμογής και τη συνολική απόδοση του πράκτορα ενισχυτικής μάθησης.

Στο πλαίσιο της ενισχυτικής μάθησης, η συνάρτηση αξίας δράσης Q(s,a) εκφράζει το αναμενόμενο αθροιστικό προεξοφλημένο όφελος που μπορεί να επιτύχει ο πράκτορας όταν βρίσκεται σε κατάσταση s, επιλέγει ενέργεια a και ακολουθεί στη συνέχεια τη μαθημένη πολιτική του. Η συνάρτηση αυτή επιτρέπει στον πράκτορα να αξιολογεί όχι μόνο το άμεσο αποτέλεσμα μιας ενέργειας, αλλά και τις μακροπρόθεσμες συνέπειες των αποφάσεών του στο χαρτοφυλάκιο.

Όπως αναλύθηκε στην Ενότητα 2.6, η εκπαίδευση της συνάρτησης αξίας δράσης στηρίζεται στο θεωρητικό πλαίσιο της εξίσωσης Bellman, το οποίο χρησιμοποιείται για τον υπολογισμό των τιμών-στόχων κατά τη διαδικασία βελτιστοποίησης του πράκτορα. Στην παρούσα εργασία, η συνάρτηση Q(s,a) δεν αποθηκεύεται σε πίνακα, αλλά προσεγγίζεται μέσω βαθύ νευρωνικού δικτύου στο πλαίσιο της αρχιτεκτονικής Dueling Deep Q-Network. Το δίκτυο εκπαιδεύεται ώστε να ελαχιστοποιεί το σφάλμα μεταξύ της τρέχουσας εκτίμησης και της τιμής-στόχου Bellman, επιτρέποντας τη σταδιακή βελτιστοποίηση της πολιτικής λήψης αποφάσεων.

Ιδιαίτερα κρίσιμος είναι ο σχεδιασμός της συνάρτησης ανταμοιβής, καθώς αυτή μεταφράζει τους στρατηγικούς στόχους της εργασίας σε μαθησιακά σήματα. Στην παρούσα εργασία, η

ανταμοιβή βασίζεται στη μεταβολή της αξίας του χαρτοφυλακίου, ενώ παράλληλα ενσωματώνει μηχανισμούς προσαρμοσμένους στον κίνδυνο, ώστε η μαθησιακή διαδικασία να λαμβάνει υπόψη όχι μόνο την επίτευξη κέρδους αλλά και τη σταθερότητα της στρατηγικής.

Η βασική μορφή της ανταμοιβής σε κάθε χρονικό βήμα ορίζεται ως:

$$r_t = V_t - V_{t-1} \qquad (3.3)$$

όπου:

r_t είναι η ανταμοιβή στο χρονικό βήμα t,

V_t είναι η συνολική αξία του χαρτοφυλακίου στο τρέχον χρονικό βήμα,

V_{t-1} είναι η συνολική αξία του χαρτοφυλακίου στο προηγούμενο χρονικό βήμα.

Η συνάρτηση αποτυπώνει τη θεμελιώδη οικονομική λογική της ανταμοιβής, σύμφωνα με την οποία ο πράκτορας επιβραβεύεται όταν οι αποφάσεις του αυξάνουν την αξία του χαρτοφυλακίου και τιμωρείται όταν οδηγούν σε απώλειες.

Σε πραγματικές συνθήκες διαπραγμάτευσης, κάθε συναλλαγή συνοδεύεται από κόστος εκτέλεσης. Για τον λόγο αυτό, η βασική συνιστώσα της ανταμοιβής μπορεί να εκφραστεί ως:

$$r_{base(t)} = r_{pnl(t)} - TC_t \qquad (3.4)$$

όπου:

$r_{base(t)}$ είναι η βασική ανταμοιβή μετά την αφαίρεση του κόστους συναλλαγών,

$r_{pnl(t)}$ είναι η συνιστώσα ανταμοιβής που προκύπτει από τη μεταβολή της αξίας του χαρτοφυλακίου, και

TC_t είναι το συνολικό κόστος συναλλαγών στο χρονικό βήμα t.

Η μορφή αυτή επιτρέπει πιο ρεαλιστική προσομοίωση της λειτουργίας των χρηματοοικονομικών αγορών, καθώς αποθαρρύνει στρατηγικές υπερβολικής συναλλακτικής δραστηριότητας που θα μπορούσαν να εμφανίζονται τεχνητά κερδοφόρες υπό συνθήκες μηδενικού κόστους.

Στην τελική υλοποίηση του συστήματος χρησιμοποιείται συνάρτηση ανταμοιβής προσαρμοσμένη στον κίνδυνο (risk-aware reward function), η οποία συνδυάζει τη βασική οικονομική απόδοση με δείκτες αξιολόγησης κινδύνου. Η συνολική ανταμοιβή ορίζεται ως:

$$r_t = a\, r_{base(t)}\, \beta\, SR_t - \gamma DD_t \qquad (3.5)$$

όπου:

r_t είναι η τελική ανταμοιβή που χρησιμοποιείται για την εκπαίδευση του πράκτορα,

$r_{base(t)}$είναι η βασική ανταμοιβή της εξίσωσης (3.4),

SR_t είναι ο δείκτης Sortino Ratio στο χρονικό βήμα t,

DD_t είναι το drawdown του χαρτοφυλακίου στο χρονικό βήμα t,

α, β και γ είναι συντελεστές στάθμισης των επιμέρους όρων της ανταμοιβής.

Στην παρούσα εργασία χρησιμοποιούνται οι τιμές:

α = 0.8

β = 0.3

γ = 0.2

Ο δείκτης Sortino Ratio χρησιμοποιείται για την αξιολόγηση της απόδοσης σε σχέση με τον καθοδικό κίνδυνο (downside risk). Σε αντίθεση με γενικότερους δείκτες μεταβλητότητας, επικεντρώνεται αποκλειστικά στις αρνητικές αποκλίσεις των αποδόσεων και συνεπώς παρέχει πιο στοχευμένη εκτίμηση του επενδυτικού κινδύνου.

Αντίστοιχα, το drawdown εκφράζει το ποσοστό υποχώρησης της αξίας του χαρτοφυλακίου από το μέγιστο επίπεδο που έχει επιτευχθεί μέχρι τη δεδομένη χρονική στιγμή. Η ενσωμάτωση του drawdown ως όρου ποινής επιτρέπει τον περιορισμό στρατηγικών που συνοδεύονται από μεγάλες προσωρινές απώλειες, ακόμη και όταν εμφανίζουν υψηλή τελική απόδοση.

Η χρήση της εξίσωσης (3.5) επιτρέπει στον πράκτορα να αναζητά στρατηγικές οι οποίες συνδυάζουν υψηλή απόδοση με αποτελεσματικότερο έλεγχο κινδύνου. Με τον τρόπο αυτό, η διαδικασία μάθησης δεν επικεντρώνεται αποκλειστικά στη μεγιστοποίηση του κέρδους, αλλά λαμβάνει υπόψη και τη σταθερότητα της επενδυτικής συμπεριφοράς.

Ιδιαίτερα σημαντικό στοιχείο της μεθοδολογίας είναι ότι η διερεύνηση της επίδρασης του δείκτη VWAP δεν πραγματοποιείται μέσω τροποποίησης της συνάρτησης ανταμοιβής, αλλά

μέσω διαφοροποίησης της αναπαράστασης της κατάστασης. Στα βασικά πειραματικά σενάρια (BASE), ο πράκτορας λειτουργεί χωρίς την πληροφορία του VWAP, ενώ στα εναλλακτικά σενάρια (VWAP), ο δείκτης ενσωματώνεται ως πρόσθετο χαρακτηριστικό εισόδου στο διάνυσμα κατάστασης. Η επιλογή αυτή διατηρεί σταθερή τη συνάρτηση ανταμοιβής μεταξύ των συγκρινόμενων εκδοχών, επιτρέποντας καθαρότερη και μεθοδολογικά αυστηρότερη αποτίμηση της πραγματικής συμβολής του VWAP στη διαδικασία λήψης αποφάσεων.

Συνολικά, η συνάρτηση αξίας δράσης και η συνάρτηση ανταμοιβής αποτελούν τον βασικό μηχανισμό μέσω του οποίου οι στρατηγικοί στόχοι της εργασίας μετατρέπονται σε μαθησιακά σήματα. Μέσω του συνδυασμού οικονομικής απόδοσης και μηχανισμών ελέγχου κινδύνου, ο πράκτορας μαθαίνει πολιτικές συναλλαγών που επιδιώκουν όχι μόνο τη μεγιστοποίηση της κερδοφορίας αλλά και τη διατήρηση πιο σταθερής και ανθεκτικής συμπεριφοράς του χαρτοφυλακίου.

3.4Ενσωμάτωση του δείκτη VWAP στην αρχιτεκτονική Dueling DQN

Ο δείκτης Volume Weighted Average Price (VWAP), όπως παρουσιάστηκε στο θεωρητικό υπόβαθρο, αποτελεί σημαντικό χρηματοοικονομικό δείκτη που συνδυάζει πληροφορία τιμής και όγκου συναλλαγών, παρέχοντας πληρέστερη εικόνα της δυναμικής της αγοράς κατά τη διάρκεια μιας χρηματιστηριακής περιόδου. Παρότι στη διεθνή βιβλιογραφία και στην πρακτική της αλγοριθμικής διαπραγμάτευσης χρησιμοποιείται κυρίως ως benchmark αξιολόγησης ποιότητας εκτέλεσης ή ως εργαλείο στρατηγικής εκτέλεσης εντολών, στην παρούσα εργασία υιοθετείται διαφορετική προσέγγιση.

Συγκεκριμένα, ο VWAP δεν χρησιμοποιείται ούτε ως μηχανισμός εκτέλεσης εντολών ούτε ως εξωτερικό benchmark ποιότητας συναλλαγών, ενώ δεν ενσωματώνεται και στη συνάρτηση ανταμοιβής του πράκτορα. Αντίθετα, αξιοποιείται αποκλειστικά ως πρόσθετο χαρακτηριστικό εισόδου (feature) στο διάνυσμα κατάστασης του πράκτορα ενισχυτικής μάθησης. Η μεθοδολογική αυτή επιλογή επιτρέπει στο σύστημα να αξιοποιεί επιπλέον πληροφορία σχετική με τη σχέση τιμής–όγκου, χωρίς να μεταβάλλεται ο βασικός μηχανισμός μάθησης ή η οικονομική λογική της reward function. Με τον τρόπο αυτό, διερευνάται με μεγαλύτερη σαφήνεια κατά πόσο η ενσωμάτωση του VWAP βελτιώνει την ποιότητα λήψης αποφάσεων.

Ο δείκτης υπολογίζεται δυναμικά σε κάθε χρονικό βήμα, αποκλειστικά με βάση τα διαθέσιμα ιστορικά δεδομένα έως τη συγκεκριμένη χρονική στιγμή. Η προσέγγιση αυτή διασφαλίζει αιτιοκρατική συνέπεια (causal consistency), δηλαδή χρήση μόνο διαθέσιμης πληροφορίας, αποφεύγει διαρροή πληροφορίας (data leakage) μέσω μη χρήσης μελλοντικών δεδομένων και εξασφαλίζει ρεαλιστική προσομοίωση πραγματικών συνθηκών αγοράς. Κατά συνέπεια, ο πράκτορας λειτουργεί σε περιβάλλον που αντανακλά με μεγαλύτερη ακρίβεια τους περιορισμούς πραγματικής εφαρμογής.

Η ενσωμάτωση του VWAP στο διάνυσμα κατάστασης (state vector) προσθέτει έναν επιπλέον πληροφοριακό άξονα στο περιβάλλον μάθησης, επιτρέποντας στον πράκτορα να αξιολογεί τη σχετική θέση της τρέχουσας τιμής σε σχέση με ένα δυναμικό σημείο αναφοράς που λαμβάνει υπόψη τη ρευστότητα της αγοράς. Έτσι, το σύστημα μπορεί να αναγνωρίζει αποκλίσεις μεταξύ τρέχουσας τιμής και μέσου σταθμισμένου επιπέδου αγοράς, να εντοπίζει πιθανές ανισορροπίες και να προσαρμόζει αναλόγως τη στρατηγική του.

Σημαντικό είναι ότι η ενσωμάτωση του VWAP δεν επιφέρει τροποποίηση στον χώρο ενεργειών, στη συνάρτηση ανταμοιβής, στη βασική δομή του Dueling DQN ή στον μηχανισμό εκπαίδευσης. Ως αποτέλεσμα, η επίδραση του δείκτη μπορεί να απομονωθεί με μεγαλύτερη ακρίβεια και να αξιολογηθεί με σαφήνεια ως προς τη συμβολή του στη μαθησιακή διαδικασία.

Για τον σκοπό αυτό, υλοποιούνται δύο συγκριτικές εκδοχές του συστήματος. Στη βασική εκδοχή (BASE), το state vector δεν περιλαμβάνει πληροφορία σχετική με τον δείκτη VWAP, ενώ στη δεύτερη εκδοχή (VWAP), το διάνυσμα κατάστασης εμπλουτίζεται με πρόσθετα χαρακτηριστικά που προκύπτουν από τον δείκτη VWAP, συμπεριλαμβανομένης της τιμής VWAP και της σχετικής απόκλισης της τρέχουσας τιμής από αυτόν (VWAP deviation). Η συγκριτική αυτή διάρθρωση επιτρέπει αυστηρότερη αξιολόγηση της πραγματικής συνεισφοράς του δείκτη, διασφαλίζοντας μεθοδολογική καθαρότητα, σταθερότητα συμπερασμάτων και αυξημένη εσωτερική εγκυρότητα.

Ένα από τα βασικά πλεονεκτήματα της προσέγγισης αυτής είναι ο σαφής διαχωρισμός μεταξύ χαρακτηριστικών εισόδου, μηχανισμού ανταμοιβής και αρχιτεκτονικής μάθησης. Ο διαχωρισμός αυτός ενισχύει την επιστημονική αξιοπιστία της πειραματικής διαδικασίας και διευκολύνει ουσιαστικά την ερμηνεία των αποτελεσμάτων, καθώς οι διαφοροποιήσεις στην απόδοση μπορούν να αποδοθούν με μεγαλύτερη βεβαιότητα στην πληροφοριακή συνεισφορά του VWAP.

Συνολικά, η ενσωμάτωση του VWAP στην αρχιτεκτονική Dueling DQN λειτουργεί ως διαδικασία εμπλουτισμού χαρακτηριστικών (feature augmentation), επιτρέποντας τη διερεύνηση του κατά πόσο η πληροφορία τιμής–όγκου μπορεί να ενισχύσει τη μαθησιακή διαδικασία και τη στρατηγική συμπεριφορά του πράκτορα, χωρίς να αλλοιώνει τις βασικές αρχές λειτουργίας του συστήματος. Με τον τρόπο αυτό, η εργασία διαφοροποιείται από τις παραδοσιακές προσεγγίσεις χρήσης του VWAP και προσφέρει ένα εναλλακτικό πλαίσιο αξιοποίησής του στο πεδίο της βαθιάς ενισχυτικής μάθησης για χρηματοοικονομικές εφαρμογές.

3.5 Προεπεξεργασία δεδομένων και τεχνικοί δείκτες

Η προεπεξεργασία των δεδομένων αποτελεί ιδιαίτερα κρίσιμο στάδιο της μεθοδολογίας της παρούσας εργασίας, καθώς επηρεάζει άμεσα τη σταθερότητα της εκπαίδευσης, την ταχύτητα σύγκλισης του νευρωνικού δικτύου και τη δυνατότητα γενίκευσης του πράκτορα ενισχυτικής μάθησης. Δεδομένου ότι η προτεινόμενη προσέγγιση βασίζεται σε ιστορικά ενδοημερήσια δεδομένα χρονικής ανάλυσης επιπέδου λεπτού (minute-level data), τα οποία χαρακτηρίζονται από υψηλό θόρυβο, σημαντική μεταβλητότητα και μη στασιμότητα, η συστηματική επεξεργασία τους είναι απαραίτητη για τη διαμόρφωση αξιόπιστου μαθησιακού περιβάλλοντος.

Τα δεδομένα που χρησιμοποιούνται προέρχονται από ιστορικές ενδοημερήσιες χρονοσειρές μετοχών του δείκτη S&P 500, διαθέσιμες μέσω της πλατφόρμας Kaggle στον σύνδεσμο: https://www.kaggle.com/datasets/nickdl/snp-500-intraday-data, και περιλαμβάνουν βασικές πληροφορίες αγοράς, όπως τιμές ανοίγματος, κλεισίματος, υψηλής και χαμηλής τιμής, καθώς και τον αντίστοιχο όγκο συναλλαγών. Η χρονική ανάλυση ανά λεπτό καθιστά το dataset ιδιαίτερα κατάλληλο για την ανάπτυξη και αξιολόγηση στρατηγικών ενδοημερήσιας διαπραγμάτευσης, καθώς επιτρέπει τη λεπτομερή αποτύπωση της μικροδομής της αγοράς.

Σε πρώτο στάδιο εφαρμόζονται διαδικασίες καθαρισμού των δεδομένων, οι οποίες περιλαμβάνουν τη διαχείριση ελλιπών τιμών, τη διόρθωση ή απομάκρυνση ακραίων ανωμαλιών, τον έλεγχο χρονικής συνέπειας και τη διασφάλιση ομοιογενούς δομής των επεισοδίων. Η διαδικασία αυτή είναι ιδιαίτερα σημαντική, καθώς ακόμη και περιορισμένος αριθμός εσφαλμένων εγγραφών μπορεί να επηρεάσει δυσανάλογα τη μαθησιακή διαδικασία σε περιβάλλοντα υψηλής συχνότητας.

Στη συνέχεια εφαρμόζεται κανονικοποίηση των δεδομένων, με στόχο τη σταθεροποίηση της εκπαιδευτικής διαδικασίας, την ταχύτερη σύγκλιση του μοντέλου, τον περιορισμό αριθμητικών ανισορροπιών μεταξύ διαφορετικών χαρακτηριστικών και τη βελτίωση της αποτελεσματικότητας των διαδικασιών βελτιστοποίησης. Η κανονικοποίηση πραγματοποιείται σε δύο διαδοχικά στάδια.

Αρχικά εφαρμόζεται τυποποίηση (standardization) μέσω του μετασχηματισμού z-score με χρήση του StandardScaler, κατά τον οποίο κάθε χαρακτηριστικό μετασχηματίζεται ώστε να έχει μηδενικό μέσο και μοναδιαία διακύμανση. Η διαδικασία αυτή συμβάλλει στη

σταθερότητα της εκπαίδευσης του νευρωνικού δικτύου και αποτρέπει την κυριαρχία χαρακτηριστικών με μεγαλύτερη αριθμητική κλίμακα. Στο στάδιο αυτό η τιμή κλεισίματος (close) δεν κανονικοποιείται, ώστε να διατηρείται η οικονομική ερμηνευσιμότητα της μεταβλητής.

Στη συνέχεια εφαρμόζεται πρόσθετη κανονικοποίηση με βάση τα στατιστικά χαρακτηριστικά του συνόλου εκπαίδευσης (μέσος όρος και τυπική απόκλιση), η οποία εφαρμόζεται ομοιόμορφα στα σύνολα εκπαίδευσης, αξιολόγησης και δοκιμής. Με τον τρόπο αυτό χρησιμοποιείται αποκλειστικά πληροφορία που προέρχεται από το σύνολο εκπαίδευσης, αποφεύγεται η διαρροή πληροφορίας (data leakage) και διασφαλίζεται η χρονική αιτιότητα της διαδικασίας.

Πέρα από τα πρωτογενή δεδομένα αγοράς, το σύστημα ενσωματώνει ένα σύνολο τεχνικών δεικτών, οι οποίοι λειτουργούν ως δευτερογενή χαρακτηριστικά υψηλότερου επιπέδου. Συγκεκριμένα, αξιοποιούνται οι δείκτες Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Bollinger Bands, Stochastic Oscillator, Schaff Trend Cycle (STC), True Range (TR) και Average True Range (ATR). Οι δείκτες αυτοί παρέχουν συμπληρωματική πληροφορία σχετικά με την ορμή, την τάση, τη μεταβλητότητα, την κυκλικότητα και πιθανές συνθήκες υπεραγοράς ή υπερπώλησης της αγοράς. Η χρήση τους δεν επιβάλλει προκαθορισμένους κανόνες συναλλαγών, αλλά λειτουργεί υποστηρικτικά, επιτρέποντας στον πράκτορα να αξιοποιεί πιο σύνθετες και χρηματοοικονομικά ερμηνεύσιμες αναπαραστάσεις της αγοράς.

Στο πλαίσιο της διαδικασίας feature engineering, αξιοποιούνται επίσης κανονικοποιημένες σχετικές αποκλίσεις που επιτρέπουν την αποτύπωση της θέσης της τρέχουσας τιμής σε σχέση με βασικά δυναμικά σημεία αναφοράς της αγοράς. Ειδικότερα, η σχετική απόκλιση από τον δείκτη VWAP ορίζεται ως:

$$d_{VWAP(t)} = \frac{Close(t) - VWAP(t)}{VWAP(t)} \qquad (3.6)$$

όπου:

Close(t) είναι η τιμή κλεισίματος στη χρονική στιγμή t,

VWAP(t) είναι η τιμή του δείκτη VWAP,

Η μεταβλητή αυτή επιτρέπει στον πράκτορα να αξιολογεί τη σχετική θέση της τρέχουσας τιμής σε σχέση με το μέσο σταθμισμένο επίπεδο τιμής–όγκου της αγοράς, παρέχοντας πληροφορία που δεν είναι άμεσα διαθέσιμη από τις πρωτογενείς μεταβλητές.

Ιδιαίτερη σημασία έχει και η δυναμική ενσωμάτωση του δείκτη VWAP, ο οποίος υπολογίζεται σε κάθε χρονικό βήμα αποκλειστικά με βάση τις διαθέσιμες πληροφορίες έως τη συγκεκριμένη χρονική στιγμή, όπως αναλύθηκε στην Ενότητα 2.8. Η προσέγγιση αυτή διατηρεί τη χρονική αιτιότητα (causal consistency) της διαδικασίας και αποφεύγει πληροφοριακή διαρροή (data leakage) μέσω χρήσης μελλοντικών δεδομένων. Με τον τρόπο αυτό, ο δείκτης λειτουργεί ως πρόσθετο πληροφοριακό χαρακτηριστικό που επιτρέπει στον πράκτορα να αξιολογεί τη σχέση μεταξύ τιμής και όγκου σε πραγματικό χρόνο.

Η τελική αναπαράσταση κατάστασης (state representation) που χρησιμοποιείται από το Dueling DQN προκύπτει από τον συνδυασμό κανονικοποιημένων χαρακτηριστικών αγοράς, τεχνικών δεικτών, μεταβλητών σχετικών με τη μεταβλητότητα της αγοράς, χαρακτηριστικών που βασίζονται στον VWAP όπου εφαρμόζεται, καθώς και μεταβλητών που περιγράφουν την κατάσταση του χαρτοφυλακίου και τη θέση του πράκτορα. Η σχεδίαση αυτή επιδιώκει ισορροπία μεταξύ πληρότητας πληροφορίας, υπολογιστικής αποδοτικότητας, περιορισμού της υπερδιάστασης και σταθερότητας της μαθησιακής διαδικασίας.

Η modular δομή της προεπεξεργασίας επιτρέπει επίσης τη δημιουργία συγκριτικών πειραματικών διαμορφώσεων, όπου η βασική εκδοχή του συστήματος λειτουργεί χωρίς χαρακτηριστικά που βασίζονται στον VWAP (BASE), ενώ η εναλλακτική εκδοχή περιλαμβάνει τόσο την τιμή VWAP όσο και τη σχετική απόκλιση από αυτήν (VWAP deviation) ως πρόσθετα χαρακτηριστικά εισόδου. Η προσέγγιση αυτή καθιστά δυνατή την αυστηρή συγκριτική αξιολόγηση της πραγματικής συνεισφοράς του δείκτη στην απόδοση του πράκτορα.

Συνολικά, η διαδικασία προεπεξεργασίας δημιουργεί ένα συνεκτικό, σταθερό και μεθοδολογικά αξιόπιστο πληροφοριακό υπόβαθρο για την εκπαίδευση του συστήματος Dueling DQN, μειώνοντας τον θόρυβο των δεδομένων, ενισχύοντας τη μαθησιακή αποδοτικότητα και επιτρέποντας την ανάπτυξη πιο αξιόπιστων και ρεαλιστικών στρατηγικών αυτόματης ενδοημερήσιας διαπραγμάτευσης.

3.6 Πλαίσιο προσομοίωσης και αξιολόγησης

Το πλαίσιο προσομοίωσης και αξιολόγησης της παρούσας εργασίας σχεδιάστηκε με στόχο τη συστηματική, πολυδιάστατη και μεθοδολογικά συνεπή αποτίμηση της απόδοσης του πράκτορα ενισχυτικής μάθησης τόσο κατά τη διάρκεια της εκπαίδευσης όσο και κατά την εφαρμογή του στα δύο πειραματικά σενάρια της εργασίας. Η αξιολόγηση δεν περιορίζεται αποκλειστικά στην τελική κερδοφορία, αλλά περιλαμβάνει συνδυασμό μαθησιακών, οικονομικών και συμπεριφορικών μετρικών, επιτρέποντας ολοκληρωμένη ανάλυση της στρατηγικής συμπεριφοράς, της σταθερότητας και της δυνατότητας προσαρμογής του συστήματος.

Το περιβάλλον προσομοίωσης οργανώνεται σε επεισόδια (episodes), στα οποία ο πράκτορας αλληλεπιδρά με ιστορικά ενδοημερήσια δεδομένα αγοράς. Η δομή των επεισοδίων διαφοροποιείται ανάλογα με το εκάστοτε πειραματικό σενάριο, με κάθε σενάριο να εξυπηρετεί διαφορετικό αλλά συμπληρωματικό ερευνητικό σκοπό.

Το πρώτο πειραματικό σενάριο, Pilot Walk-Forward Scenario (6-Day Dataset), βασίζεται σε έξι διαδοχικές χρηματιστηριακές συνεδριάσεις και λειτουργεί ως ελεγχόμενο πλαίσιο αρχικής διερεύνησης της λειτουργικότητας, της στρατηγικής συμπεριφοράς και της βασικής αποδοτικότητας του προτεινόμενου συστήματος. Στο σενάριο αυτό υλοποιείται ένα επεισόδιο, το οποίο περιλαμβάνει πέντε διαδοχικούς κύκλους εκπαίδευσης και συναλλαγών (Train→Trade), επιτρέποντας την άμεση παρακολούθηση της μαθησιακής διαδικασίας και την αρχική αξιολόγηση της επίδρασης της ενσωμάτωσης του δείκτη VWAP στο διάνυσμα κατάστασης.

Αντίστοιχα, το δεύτερο πειραματικό σενάριο, Extended Walk-Forward Scenario (111-Day Dataset), αξιοποιεί εκτενέστερο σύνολο δεδομένων αποτελούμενο από 111 διαδοχικές χρηματιστηριακές συνεδριάσεις. Η διαδικασία οργανώνεται σε διαδοχικά και ανεξάρτητα επεισόδια, καθένα από τα οποία ξεκινά με νέο χαρτοφυλάκιο αρχικής αξίας 100.000 δολαρίων και περιλαμβάνει πέντε διαδοχικούς κύκλους Train→Trade. Σε κάθε κύκλο ο πράκτορας εκπαιδεύεται χρησιμοποιώντας τα δεδομένα μίας συνεδρίασης και εφαρμόζει την εκμαθημένη πολιτική στην αμέσως επόμενη συνεδρία χωρίς περαιτέρω ενημέρωση των παραμέτρων του μοντέλου. Η δομή αυτή επιτρέπει την αξιολόγηση της συμπεριφοράς του πράκτορα σε μεγάλο

χρονικό εύρος και σε διαφορετικές συνθήκες αγοράς, διατηρώντας παράλληλα τη φυσική χρονική ακολουθία των δεδομένων.

Συνεπώς, τα δύο σενάρια λειτουργούν συμπληρωματικά, προσφέροντας τόσο αναλυτική αρχική αξιολόγηση όσο και εκτενέστερη πειραματική επιβεβαίωση της προτεινόμενης μεθοδολογίας.

Κατά τη διάρκεια της εκπαίδευσης, βασική μετρική αξιολόγησης αποτελεί η αθροιστική ανταμοιβή (cumulative reward), η οποία αποτυπώνει το συνολικό μαθησιακό όφελος του πράκτορα σύμφωνα με τη συνάρτηση ανταμοιβής. Η παρακολούθηση της εξέλιξής της επιτρέπει την εκτίμηση της διαδικασίας σύγκλισης, την αναγνώριση πιθανών προβλημάτων αστάθειας και τη συνολική αποτίμηση της βελτίωσης της μαθημένης πολιτικής.

Παράλληλα, καταγράφεται η εξέλιξη της αξίας του χαρτοφυλακίου (portfolio value), η οποία παρέχει άμεση οικονομική αποτίμηση της συμπεριφοράς του μοντέλου, καθώς συνδέεται με το αποτέλεσμα των διαδοχικών επενδυτικών αποφάσεων. Η μετρική αυτή διευκολύνει τη σύνδεση της μαθησιακής διαδικασίας με την αποτελεσματικότητα της στρατηγικής συναλλαγών.

Επιπλέον, αξιολογούνται συμπεριφορικές μετρικές, όπως ο αριθμός συναλλαγών, η κατανομή των ενεργειών αγοράς (long) και πώλησης (short), η συνολική συναλλακτική δραστηριότητα, καθώς και λειτουργικές μετρικές που αφορούν το κόστος συναλλαγών και τη διαχείριση θέσεων. Οι δείκτες αυτοί επιτρέπουν βαθύτερη κατανόηση της στρατηγικής που αναπτύσσει ο πράκτορας και συμβάλλουν στον εντοπισμό πιθανών μη ρεαλιστικών συμπεριφορών.

Ιδιαίτερη σημασία αποδίδεται στη συγκριτική αξιολόγηση μεταξύ των εκδοχών BASE και VWAP. Η σύγκριση πραγματοποιείται υπό απολύτως ίδιες συνθήκες αρχιτεκτονικής, reward function, trading environment και διαδικασίας εκπαίδευσης, ώστε η μοναδική διαφοροποίηση να αφορά την παρουσία ή μη των χαρακτηριστικών VWAP στο state vector. Με τον τρόπο αυτό απομονώνεται η πληροφοριακή συνεισφορά του VWAP και διασφαλίζεται η μεθοδολογική εγκυρότητα της σύγκρισης.

Οι βασικές τελικές μετρικές αξιολόγησης περιλαμβάνουν την τελική αξία του χαρτοφυλακίου (PnL), την εξέλιξη της αξίας του χαρτοφυλακίου, τη συναλλακτική δραστηριότητα, καθώς και τη συνέπεια της συμπεριφοράς του πράκτορα μεταξύ διαφορετικών επεισοδίων και διαφορετικών χρονικών περιόδων της αγοράς. Η συνδυαστική αξιοποίηση των μετρικών

αυτών επιτρέπει πολυδιάστατη αξιολόγηση, υπερβαίνοντας τη μονοδιάστατη προσέγγιση της απλής τελικής κερδοφορίας.

Συνολικά, το πλαίσιο προσομοίωσης και αξιολόγησης της παρούσας εργασίας εξασφαλίζει υψηλό επίπεδο μεθοδολογικής αυστηρότητας, συγκρισιμότητας και επιστημονικής αξιοπιστίας. Μέσω της συνδυασμένης ανάλυσης μαθησιακών, οικονομικών και συμπεριφορικών δεικτών, επιτρέπει την ολοκληρωμένη διερεύνηση της λειτουργικότητας, της στρατηγικής ποιότητας και της πρακτικής χρησιμότητας του προτεινόμενου συστήματος αυτόματης διαπραγμάτευσης.

Οι παραπάνω μετρικές χρησιμοποιούνται κυρίως για την παρακολούθηση της διαδικασίας εκπαίδευσης και για τη συγκριτική ανάλυση της συμπεριφοράς του πράκτορα μεταξύ των δύο πειραματικών σεναρίων. Η τελική αξιολόγηση βασίζεται κυρίως στην απόδοση που επιτυγχάνει ο πράκτορας κατά τις φάσεις συναλλαγών (Trade), όπως αυτή αποτυπώνεται στην τελική αξία του χαρτοφυλακίου, στην εξέλιξη της επενδυτικής στρατηγικής και στη συνολική συναλλακτική συμπεριφορά.

3.7 Παράμετροι εκπαίδευσης και διαδικασία σύγκλισης

Η διαδικασία εκπαίδευσης του πράκτορα ενισχυτικής μάθησης βασίζεται σε ένα σύνολο υπερπαραμέτρων (hyperparameters), οι οποίες επηρεάζουν καθοριστικά τη σταθερότητα, την ταχύτητα προσαρμογής και τη δυνατότητα σύγκλισης του μοντέλου. Οι βασικές αυτές παράμετροι περιλαμβάνουν τον συντελεστή μάθησης (learning rate), τον συντελεστή προεξόφλησης (discount factor), το μέγεθος της μνήμης εμπειρικής επανάληψης (replay memory), το μέγεθος των παρτίδων εκπαίδευσης (batch size), καθώς και τη στρατηγική εξερεύνησης. Η επιλογή τους βασίζεται τόσο σε καθιερωμένες πρακτικές της βιβλιογραφίας όσο και σε πειραματική διαδικασία αρχικής παραμετροποίησης, με στόχο την επίτευξη σταθερής και συνεπούς εκπαίδευσης του πράκτορα, εξασφαλίζοντας ικανοποιητική σύγκλιση του μοντέλου χωρίς υπερβολική αύξηση του υπολογιστικού κόστους. Οι συγκεκριμένες τιμές των υπερπαραμέτρων που χρησιμοποιούνται στην υλοποίηση παρουσιάζονται αναλυτικά στο Κεφάλαιο 4, ώστε να διασφαλίζεται η αναπαραγωγιμότητα της πειραματικής διαδικασίας.

Η εκπαίδευση οργανώνεται σε διακριτά επεισόδια, εντός των οποίων ο πράκτορας αλληλεπιδρά με το περιβάλλον διαπραγμάτευσης, συλλέγοντας εμπειρίες και προσαρμόζοντας σταδιακά τη συνάρτηση αξίας δράσης (Q(s,a)). Η ακριβής δομή των επεισοδίων διαφοροποιείται μεταξύ των δύο πειραματικών σεναρίων της εργασίας. Στο Pilot Walk-Forward Scenario (6-Day Dataset) υλοποιείται ένα επεισόδιο αποτελούμενο από πέντε διαδοχικούς κύκλους Train→Trade, ενώ στο Extended Walk-Forward Scenario (111-Day Dataset) η διαδικασία επαναλαμβάνεται σε πολλαπλά ανεξάρτητα επεισόδια, καθένα από τα οποία περιλαμβάνει επίσης πέντε διαδοχικούς κύκλους Train→Trade και ξεκινά με νέο χαρτοφυλάκιο αρχικής αξίας 100.000 δολαρίων. Η σωρευτική ανταμοιβή χρησιμοποιείται ως βασικός εσωτερικός δείκτης παρακολούθησης της μαθησιακής προόδου και της σταδιακής σταθεροποίησης της στρατηγικής. Δεδομένου ότι η συνάρτηση ανταμοιβής ενσωματώνει τόσο στοιχεία απόδοσης όσο και μηχανισμούς ελέγχου κινδύνου, η εξέλιξη της σωρευτικής ανταμοιβής παρέχει συνολική εικόνα της μαθησιακής προόδου του πράκτορα.

Η πολιτική εξερεύνησης υλοποιείται μέσω στρατηγικής ε-greedy, κατά την οποία με πιθανότητα ε επιλέγεται τυχαία ενέργεια από τον χώρο ενεργειών, ενώ με πιθανότητα 1−ε επιλέγεται η ενέργεια με τη μέγιστη εκτιμώμενη τιμή της συνάρτησης Q(s,a).

Στα αρχικά στάδια της εκπαίδευσης, υψηλότερες τιμές του ε επιτρέπουν ευρύτερη εξερεύνηση του χώρου ενεργειών, ενώ η σταδιακή απομείωσή του οδηγεί προοδευτικά σε μεγαλύτερη εκμετάλλευση της αποκτηθείσας γνώσης. Η διαδικασία αυτή είναι απαραίτητη για την αποφυγή πρόωρης σύγκλισης σε υποβέλτιστες πολιτικές.

Για τη βελτίωση της σταθερότητας της μάθησης χρησιμοποιείται μηχανισμός επανάληψης εμπειριών (experience replay), όπου οι μεταβάσεις εμπειρίας του πράκτορα αποθηκεύονται προσωρινά και επαναδειγματοληπτούνται τυχαία κατά τη διαδικασία ενημέρωσης του νευρωνικού δικτύου. Η προσέγγιση αυτή περιορίζει τη χρονική συσχέτιση των διαδοχικών εμπειριών, αυξάνει τη στατιστική αποδοτικότητα και βελτιώνει τη συνολική σταθερότητα της εκπαίδευσης.

Παράλληλα, χρησιμοποιείται target network, το οποίο ενημερώνεται περιοδικά και χρησιμοποιείται για τον υπολογισμό σταθερότερων τιμών-στόχων κατά τη διαδικασία βελτιστοποίησης. Ο διαχωρισμός αυτός μεταξύ κύριου δικτύου και δικτύου στόχου συμβάλλει στον περιορισμό ασταθών διακυμάνσεων και στην ομαλότερη σύγκλιση της μαθησιακής διαδικασίας.

Η διαδικασία σύγκλισης παρακολουθείται κυρίως μέσω της σταθεροποίησης της σωρευτικής ανταμοιβής, της εξέλιξης των συναρτήσεων απώλειας (loss functions) και της συνολικής συμπεριφοράς των εκπαιδευτικών μετρικών. Η παρακολούθηση αυτή επιτρέπει τον εντοπισμό προβλημάτων όπως ασταθής μάθηση, υπερεκπαίδευση ή ανεπαρκής εξερεύνηση, πριν από την τελική αξιολόγηση του συστήματος.

Η πειραματική διαδικασία ξεκινά με το Pilot Walk-Forward Scenario (6-Day Dataset), το οποίο χρησιμοποιείται για την επαλήθευση της ορθότητας της υλοποίησης, της λειτουργικότητας του περιβάλλοντος και της βασικής συμπεριφοράς του πράκτορα. Στη συνέχεια εφαρμόζεται το Extended Walk-Forward Scenario (111-Day Dataset), όπου η εκπαίδευση και η αξιολόγηση επαναλαμβάνονται σε μεγαλύτερο χρονικό εύρος και σε πολλαπλά ανεξάρτητα επεισόδια, επιτρέποντας την αποτίμηση της σταθερότητας, της επαναληψιμότητας και της δυνατότητας προσαρμογής της μαθημένης πολιτικής σε διαφορετικές συνθήκες αγοράς.

Η διαφοροποίηση των χαρακτηριστικών εισόδου μεταξύ των επιμέρους πειραματικών εκδοχών του συστήματος επιτρέπει τη διερεύνηση της επίδρασης πρόσθετων πληροφοριακών μεταβλητών στη μαθησιακή διαδικασία, χωρίς να τροποποιούνται οι βασικοί μηχανισμοί εκπαίδευσης. Με τον τρόπο αυτό, καθίσταται δυνατή η σαφέστερη αποτίμηση της επίδρασης επιμέρους σχεδιαστικών επιλογών στη σταθερότητα και την αποδοτικότητα της μάθησης.

Συνολικά, οι παράμετροι εκπαίδευσης και η διαδικασία σύγκλισης έχουν σχεδιαστεί ώστε να εξασφαλίζουν υψηλό επίπεδο πειραματικού ελέγχου, σταθερότητας και μεθοδολογικής συνέπειας, διασφαλίζοντας την αξιοπιστία και την αναπαραγωγιμότητα των αποτελεσμάτων. Η προσέγγιση αυτή ενισχύει την αξιοπιστία των αποτελεσμάτων και δημιουργεί ισχυρό υπόβαθρο για την τεχνική υλοποίηση και την αναλυτική παρουσίαση των πειραματικών αποτελεσμάτων στα επόμενα κεφάλαια.

4. Υλοποίηση του συστήματος ενισχυτικής μάθησης

4.1 Αρχιτεκτονική του συστήματος

Η υλοποίηση του προτεινόμενου συστήματος ενισχυτικής μάθησης βασίζεται σε μια αρθρωτή και λειτουργικά οργανωμένη αρχιτεκτονική, η οποία επιτρέπει την αποτελεσματική ενσωμάτωση ιστορικών δεδομένων χρηματοοικονομικής αγοράς, διαδικασιών προεπεξεργασίας, περιβάλλοντος προσομοίωσης συναλλαγών και πράκτορα βαθιάς ενισχυτικής μάθησης. Η συνολική σχεδίαση ακολουθεί τις βασικές αρχές του πλαισίου Reinforcement Learning, όπου ένας πράκτορας αλληλεπιδρά διαδοχικά με ένα περιβάλλον συναλλαγών, λαμβάνοντας αποφάσεις και προσαρμόζοντας τη στρατηγική του βάσει της ανταμοιβής που προκύπτει από την επίδραση των ενεργειών του στην απόδοση του χαρτοφυλακίου και από μηχανισμούς αξιολόγησης κινδύνου που ενσωματώνονται στη διαδικασία μάθησης.

Το σύστημα δεν περιορίζεται σε μια διαδικασία πρόβλεψης μελλοντικών τιμών, αλλά λειτουργεί ως ολοκληρωμένο πλαίσιο αυτόνομης λήψης αποφάσεων, όπου ο πράκτορας επιδιώκει τη βελτιστοποίηση της συνολικής στρατηγικής συναλλαγών. Η προσέγγιση αυτή επιτρέπει τη μελέτη όχι μόνο της προβλεπτικής ικανότητας του μοντέλου αλλά κυρίως της πρακτικής του αποδοτικότητας ως μηχανισμού αλγοριθμικής διαπραγμάτευσης.

Η συνολική λειτουργία του συστήματος οργανώνεται γύρω από τη βασική σχέση μεταξύ περιβάλλοντος και πράκτορα. Τα δεδομένα αγοράς, οι τεχνικοί δείκτες και οι διαδικασίες προεπεξεργασίας λειτουργούν ως υποστηρικτικές συνιστώσες που διαμορφώνουν την αναπαράσταση της κατάστασης του περιβάλλοντος, χωρίς να αποτελούν αυτοτελή υποσυστήματα ενισχυτικής μάθησης. Ο πυρήνας του συστήματος είναι η συνεχής αλληλεπίδραση του πράκτορα με το trading environment, εντός του οποίου πραγματοποιείται η μαθησιακή διαδικασία.

Σε κάθε χρονικό βήμα t, το περιβάλλον παρέχει στον πράκτορα ένα διάνυσμα κατάστασης s_t, το οποίο περιλαμβάνει πληροφορίες αγοράς, μεταβολές όγκου συναλλαγών, τεχνικούς δείκτες, χαρακτηριστικά μεταβλητότητας, μεταβλητές χαρτοφυλακίου και θέσης, καθώς και χαρακτηριστικά που βασίζονται στον δείκτη VWAP όπου εφαρμόζεται. Ο πράκτορας επιλέγει μία ενέργεια a_t από τον διακριτό χώρο ενεργειών (αγορά, πώληση ή διακράτηση), η οποία

εφαρμόζεται στο περιβάλλον. Το περιβάλλον επιστρέφει την επόμενη κατάσταση s_{t+1} και την αντίστοιχη ανταμοιβή r_t, επιτρέποντας τη συνεχή ενημέρωση της συνάρτησης αξίας δράσης.

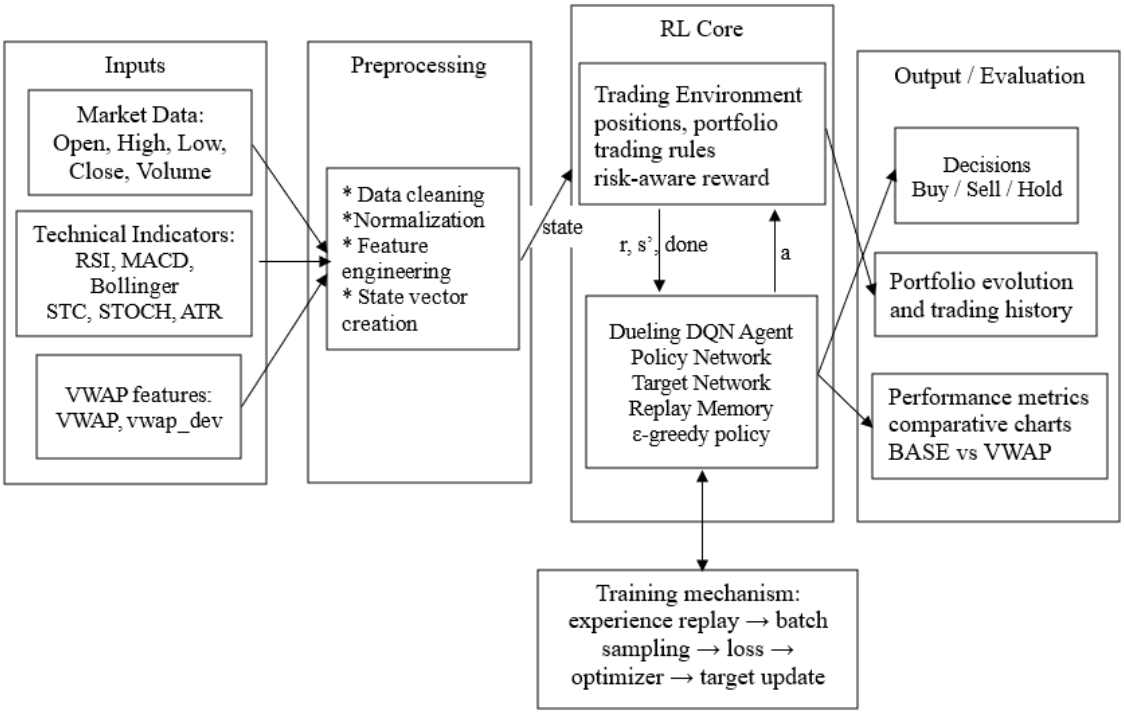
Η αρχιτεκτονική του πράκτορα βασίζεται στο Dueling Deep Q-Network (Dueling DQN), το οποίο περιλαμβάνει policy network, target network, replay memory και στρατηγική εξερεύνησης τύπου ε-greedy. Μέσω αυτής της δομής, το σύστημα επιτυγχάνει σταθερότερη εκπαίδευση, βελτιωμένη σύγκλιση και αποτελεσματικότερη εκτίμηση των αναμενόμενων αποδόσεων διαφορετικών ενεργειών.

Η χρήση της αρχιτεκτονικής Dueling DQN επιτρέπει τον διαχωρισμό της συνολικής συνάρτησης αξίας δράσης Q(s,a) στις συνιστώσες αξίας κατάστασης V(s) και πλεονεκτήματος δράσης A(s,a), ενισχύοντας τη μαθησιακή αποδοτικότητα σε περιβάλλοντα όπου διαφορετικές ενέργειες συχνά οδηγούν σε παρόμοια αποτελέσματα, όπως συμβαίνει στις χρηματοοικονομικές αγορές.

Παράλληλα, η υλοποίηση διατηρεί αρθρωτό (modular) χαρακτήρα, καθώς οι επιμέρους συνιστώσες — preprocessing pipeline, trading environment, agent, καθώς και τα υποσυστήματα εκπαίδευσης και αξιολόγησης — οργανώνονται με τρόπο που επιτρέπει ανεξάρτητες τροποποιήσεις, συγκριτικές δοκιμές και πειραματικές επεκτάσεις χωρίς ανασχεδιασμό του συνολικού συστήματος.

Η συγκεκριμένη αρχιτεκτονική υποστηρίζει δύο βασικές πειραματικές εκδοχές: τη βασική έκδοση BASE και την έκδοση VWAP, όπου η μόνη ουσιαστική διαφοροποίηση αφορά την ενσωμάτωση πρόσθετων χαρακτηριστικών που βασίζονται στον δείκτη VWAP, συμπεριλαμβανομένης της τιμής VWAP και της σχετικής απόκλισης της τρέχουσας τιμής από αυτόν (VWAP deviation), στο state vector. Με τον τρόπο αυτό διασφαλίζεται ότι οι συγκριτικές αξιολογήσεις αποδίδονται με σαφήνεια στην πληροφοριακή συμβολή του δείκτη και όχι σε διαφοροποιήσεις της βασικής δομής του συστήματος.

Συνολικά, η αρχιτεκτονική του συστήματος παρέχει ένα συνεκτικό, επεκτάσιμο και επιστημονικά τεκμηριωμένο πλαίσιο για την ανάπτυξη, εκπαίδευση και αξιολόγηση προηγμένων στρατηγικών βαθιάς ενισχυτικής μάθησης σε συνθήκες ενδοημερήσιας αλγοριθμικής διαπραγμάτευσης.



Σχήμα 1 Συνολική αρχιτεκτονική του συστήματος ενισχυτικής μάθησης

4.1.1 Συνολική δομή του συστήματος

Η συνολική δομή του προτεινόμενου συστήματος οργανώνεται ως μια διαδοχική και επαναληπτική διαδικασία ροής πληροφορίας, μέσω της οποίας τα ιστορικά δεδομένα χρηματοοικονομικής αγοράς μετασχηματίζονται σε κατάλληλη μορφή εισόδου για τον πράκτορα ενισχυτικής μάθησης και αξιοποιούνται για τη λήψη και βελτιστοποίηση αποφάσεων συναλλαγών. Η δομή αυτή επιτρέπει τη λειτουργική σύνδεση μεταξύ των εισόδων δεδομένων, του περιβάλλοντος συναλλαγών, του πράκτορα και των μηχανισμών εκπαίδευσης, διαμορφώνοντας ένα ολοκληρωμένο πλαίσιο αλγοριθμικής διαπραγμάτευσης.

Η διαδικασία εκκινεί από την εισαγωγή πρωτογενών δεδομένων αγοράς, τα οποία περιλαμβάνουν τιμές ανοίγματος, υψηλής, χαμηλής και κλεισίματος, όγκους συναλλαγών, καθώς και τους απαραίτητους τεχνικούς δείκτες. Στην εκδοχή VWAP, στα δεδομένα αυτά ενσωματώνονται επιπλέον χαρακτηριστικά που βασίζονται στον δείκτη VWAP, συμπεριλαμβανομένης της τιμής VWAP και της σχετικής απόκλισης της τρέχουσας τιμής από αυτόν (VWAP deviation), όπως έχει ήδη αναλυθεί στο θεωρητικό πλαίσιο. Τα δεδομένα αυτά

υποβάλλονται σε διαδικασίες καθαρισμού, κανονικοποίησης και εξαγωγής χαρακτηριστικών, προκειμένου να διαμορφωθεί μια σταθερή και κατάλληλη μαθησιακή αναπαράσταση.

Το περιβάλλον συναλλαγών αξιοποιεί τα προεπεξεργασμένα δεδομένα για τη δημιουργία του διανύσματος κατάστασης του πράκτορα, το οποίο περιγράφει την τρέχουσα εικόνα της αγοράς, τη θέση του χαρτοφυλακίου και τις σχετικές χρηματοοικονομικές μεταβλητές κάθε χρονικού βήματος. Το διάνυσμα κατάστασης (state vector) αποτελεί την κύρια είσοδο προς τον πράκτορα Dueling DQN και λειτουργεί ως πληροφοριακή βάση για τη διαδικασία λήψης αποφάσεων.

Ο πράκτορας επεξεργάζεται την κατάσταση μέσω του policy network και επιλέγει μία ενέργεια από τον διακριτό χώρο ενεργειών, δηλαδή αγορά, πώληση ή διακράτηση. Η ενέργεια εφαρμόζεται στο περιβάλλον συναλλαγών, το οποίο ενημερώνει την κατάσταση του χαρτοφυλακίου, υπολογίζει τη νέα αξία του και υπολογίζει και επιστρέφει την αντίστοιχη ανταμοιβή. Η διαδικασία αυτή επαναλαμβάνεται διαδοχικά σε κάθε χρονικό βήμα, συγκροτώντας τον βασικό κύκλο λειτουργίας του συστήματος.

Παράλληλα, κάθε εμπειρία αλληλεπίδρασης του πράκτορα με το περιβάλλον αποθηκεύεται στη replay memory, επιτρέποντας τη μεταγενέστερη δειγματοληψία και επαναχρησιμοποίησή της κατά την εκπαιδευτική διαδικασία. Το policy network ενημερώνεται συνεχώς μέσω διαδικασιών βελτιστοποίησης, ενώ το target network χρησιμοποιείται για τη σταθεροποίηση των τιμών-στόχων, περιορίζοντας την αστάθεια της μάθησης.

Η συνολική δομή του συστήματος περιλαμβάνει επίσης διακριτούς μηχανισμούς αξιολόγησης, μέσω των οποίων καταγράφονται η εξέλιξη της αξίας του χαρτοφυλακίου, το ιστορικό συναλλαγών, οι σωρευτικές ανταμοιβές, οι δείκτες Profit and Loss (PnL), οι μετρικές κινδύνου που χρησιμοποιούνται στο σύστημα, καθώς και οι συγκριτικές επιδόσεις μεταξύ των πειραματικών εκδοχών BASE και VWAP. Η δυνατότητα αυτή επιτρέπει την αυστηρή αξιολόγηση της μαθησιακής συμπεριφοράς και της χρηματοοικονομικής αποδοτικότητας του συστήματος.

Η modular σχεδίαση της συνολικής δομής διασφαλίζει σαφή διαχωρισμό λειτουργιών, αυξημένη επεκτασιμότητα και δυνατότητα ανεξάρτητης τροποποίησης επιμέρους συνιστωσών. Επιπλέον, διευκολύνει τη σαφή αντιστοίχιση μεταξύ θεωρητικού πλαισίου και πραγματικής τεχνικής υλοποίησης, ενισχύοντας τη μεθοδολογική συνέπεια και την επιστημονική τεκμηρίωση της εργασίας.

4.1.2 Βασικά δομικά στοιχεία της υλοποίησης

Η πρακτική υλοποίηση του προτεινόμενου συστήματος βασίζεται σε ένα σύνολο διακριτών αλλά λειτουργικά αλληλεξαρτώμενων υποσυστημάτων, τα οποία συνεργάζονται για να υποστηρίξουν το σύνολο της διαδικασίας από την εισαγωγή και επεξεργασία των δεδομένων έως την εκπαίδευση, αξιολόγηση και ανάλυση της συμπεριφοράς του πράκτορα ενισχυτικής μάθησης. Η υιοθέτηση αρθρωτής (modular) σχεδίασης αποτελεί κρίσιμη επιλογή, καθώς επιτρέπει σαφή διαχωρισμό ρόλων, ευκολότερη τεχνική διαχείριση, δυνατότητα ανεξάρτητης βελτίωσης επιμέρους συνιστωσών και υψηλότερη αναπαραγωγιμότητα των πειραματικών διαδικασιών.

Το πρώτο λειτουργικό υποσύστημα αφορά την εισαγωγή και διαχείριση δεδομένων, το οποίο είναι υπεύθυνο για τη φόρτωση των ιστορικών χρηματοοικονομικών δεδομένων, την επιλογή των εξεταζόμενων μετοχών, τον χρονικό διαχωρισμό των δεδομένων σε περιόδους εκπαίδευσης, αξιολόγησης και τελικού ελέγχου, καθώς και την οργάνωση των απαιτούμενων rolling-window ή walk-forward πειραματικών διατάξεων. Η μονάδα αυτή αποτελεί τη βασική πηγή τροφοδότησης του συστήματος με συνεπή και κατάλληλα οργανωμένη πληροφορία.

Ακολουθεί το υποσύστημα προεπεξεργασίας και εξαγωγής χαρακτηριστικών, το οποίο αναλαμβάνει τον καθαρισμό των δεδομένων, τη διαχείριση ελλιπών ή ακραίων τιμών, την κανονικοποίηση, την αριθμητική σταθεροποίηση και τον υπολογισμό τεχνικών δεικτών (RSI, MACD, Bollinger Bands, STC, Stochastic Oscillator, TR και ATR), καθώς και πρόσθετων χαρακτηριστικών που βασίζονται στον δείκτη VWAP, συμπεριλαμβανομένης της τιμής VWAP και της σχετικής απόκλισης της τρέχουσας τιμής από αυτόν (VWAP deviation). Στόχος του είναι η μετατροπή των πρωτογενών δεδομένων αγοράς σε συνεκτική μαθησιακή αναπαράσταση κατάστασης, κατάλληλη για χρήση από το νευρωνικό δίκτυο του πράκτορα.

Το περιβάλλον συναλλαγών αποτελεί τον κεντρικό μηχανισμό αλληλεπίδρασης μεταξύ πράκτορα και αγοράς. Είναι υπεύθυνο για την παρακολούθηση της χρονικής εξέλιξης των επεισοδίων, τη διαχείριση ανοικτών θέσεων, την εφαρμογή ενεργειών αγοράς, πώλησης ή διακράτησης, την ενημέρωση της αξίας του χαρτοφυλακίου, τον υπολογισμό της ανταμοιβής, η οποία συνδυάζει στοιχεία απόδοσης και διαχείρισης κινδύνου, και τη μετάβαση στην επόμενη κατάσταση. Μέσω του υποσυστήματος αυτού επιτυγχάνεται η πρακτική προσομοίωση της λειτουργίας πραγματικού περιβάλλοντος διαπραγμάτευσης.

Ο πράκτορας Dueling Deep Q-Network συνιστά τον βασικό πυρήνα μάθησης του συστήματος. Περιλαμβάνει το κύριο policy network, το target network, τους μηχανισμούς επιλογής ενεργειών, τη διασύνδεση με τη μνήμη εμπειριών και τις διαδικασίες βελτιστοποίησης των παραμέτρων του νευρωνικού δικτύου. Η συγκεκριμένη αρχιτεκτονική επιτρέπει πιο αποδοτική εκτίμηση των στρατηγικών επιλογών μέσω του διαχωρισμού μεταξύ αξίας κατάστασης και πλεονεκτήματος ενεργειών.

Η μνήμη εμπειριών (replay memory) λειτουργεί ως κρίσιμο υποσύστημα σταθεροποίησης της μάθησης, αποθηκεύοντας εμπειρίες της μορφής

$$(s, \mathrm{a}, r, s', done) \qquad (4.1)$$

όπου:

s: τρέχουσα κατάσταση

a: επιλεγμένη ενέργεια

r: ανταμοιβή

s′: επόμενη κατάσταση

done: ένδειξη ολοκλήρωσης επεισοδίου

ώστε να επιτρέπεται τυχαία επαναδειγματοληψία κατά την εκπαίδευση. Με τον τρόπο αυτό μειώνονται οι χρονικές συσχετίσεις μεταξύ διαδοχικών εμπειριών και ενισχύεται η στατιστική αποδοτικότητα της μαθησιακής διαδικασίας.

Η συνολική εκπαιδευτική διαδικασία διαχειρίζεται από εξειδικευμένη μονάδα εκπαίδευσης (training engine), η οποία οργανώνει την επεισοδιακή εκπαίδευση, τη δειγματοληψία παρτίδων εκπαίδευσης, τον υπολογισμό της συνάρτησης απώλειας, την εφαρμογή gradient descent, την ενημέρωση του policy network, τον συγχρονισμό του target network και τη σταδιακή απομείωση του παράγοντα εξερεύνησης. Η μονάδα αυτή εξασφαλίζει ελεγχόμενη και σταθερή προσαρμογή του πράκτορα.

Η τελική αποτίμηση πραγματοποιείται μέσω ανεξάρτητου evaluation engine, το οποίο εφαρμόζει την εκπαιδευμένη πολιτική σε δεδομένα αξιολόγησης ή ελέγχου, καταγράφει την απόδοση του χαρτοφυλακίου, υπολογίζει χρηματοοικονομικές μετρικές απόδοσης και κινδύνου, καταγράφει την εξέλιξη του χαρτοφυλακίου και υποστηρίζει τη συγκριτική

αξιολόγηση μεταξύ διαφορετικών πειραματικών σεναρίων. Η ανεξαρτησία της διαδικασίας αξιολόγησης ενισχύει την αντικειμενικότητα των αποτελεσμάτων.

Τέλος, το υποσύστημα οπτικοποίησης (visualization) και αναφοράς αποτελεσμάτων επιτρέπει την παραγωγή διαγραμμάτων, ιστορικών συναλλαγών, καμπυλών εξέλιξης χαρτοφυλακίου, ιστορικών συναλλαγών, συγκριτικών διαγραμμάτων BASE και VWAP και στατιστικών αναλύσεων, συμβάλλοντας ουσιαστικά στην ερμηνεία της συμπεριφοράς του πράκτορα και στην ακαδημαϊκή τεκμηρίωση της εργασίας.

Συνολικά, τα δομικά στοιχεία της υλοποίησης συγκροτούν ένα ολοκληρωμένο, τεχνικά συνεκτικό και επιστημονικά αξιόπιστο σύστημα βαθιάς ενισχυτικής μάθησης, το οποίο υποστηρίζει τη συστηματική ανάπτυξη, εκπαίδευση και συγκριτική αξιολόγηση στρατηγικών αλγοριθμικής διαπραγμάτευσης.

4.1.3 Ροή δεδομένων και αλληλεπίδραση των υποσυστημάτων

Η λειτουργία του προτεινόμενου συστήματος βασίζεται σε μια διαδοχική και επαναληπτική ροή δεδομένων, η οποία ξεκινά από την εισαγωγή ιστορικών χρηματοοικονομικών δεδομένων και καταλήγει στη λήψη αποφάσεων συναλλαγών, στην ενημέρωση του χαρτοφυλακίου και στην αξιολόγηση της συνολικής απόδοσης του πράκτορα. Η ροή αυτή αποτελεί την πρακτική υλοποίηση της αλληλεπίδρασης μεταξύ πράκτορα και περιβάλλοντος και συνιστά τον βασικό λειτουργικό μηχανισμό του συστήματος ενισχυτικής μάθησης.

Στο πρώτο στάδιο, το σύστημα λαμβάνει ως είσοδο ιστορικά δεδομένα αγοράς, τα οποία περιλαμβάνουν τιμές ανοίγματος, υψηλής, χαμηλής και κλεισίματος, όγκους συναλλαγών και τεχνικά χαρακτηριστικά. Στα σενάρια όπου εξετάζεται η ενσωμάτωση του VWAP, το σύνολο των χαρακτηριστικών εμπλουτίζεται με πρόσθετα χαρακτηριστικά που βασίζονται στον δείκτη VWAP, συμπεριλαμβανομένης της τιμής VWAP και της σχετικής απόκλισης της τρέχουσας τιμής από αυτόν (VWAP deviation), χωρίς να τροποποιείται ο βασικός μηχανισμός λήψης αποφάσεων ή η συνάρτηση ανταμοιβής.

Τα δεδομένα αυτά υποβάλλονται σε διαδικασίες καθαρισμού, κανονικοποίησης και μετασχηματισμού, προκειμένου να διαμορφωθούν σε αριθμητικά χαρακτηριστικά κατάλληλα για χρήση από το νευρωνικό δίκτυο. Το αποτέλεσμα αυτής της διαδικασίας είναι η δημιουργία

του διανύσματος κατάστασης, το οποίο αναπαριστά τη συνολική εικόνα της αγοράς και του χαρτοφυλακίου σε κάθε χρονικό βήμα.

Η κατάσταση αποστέλλεται στον πράκτορα Dueling DQN, ο οποίος επεξεργάζεται το state vector μέσω του policy network και επιλέγει μία ενέργεια από τον προκαθορισμένο διακριτό χώρο ενεργειών, δηλαδή αγορά, πώληση ή διακράτηση. Η επιλογή αυτή βασίζεται τόσο στη μαθημένη πολιτική όσο και στη στρατηγική εξερεύνησης που εφαρμόζεται κατά την εκπαιδευτική διαδικασία.

Η επιλεγμένη ενέργεια εφαρμόζεται στο περιβάλλον συναλλαγών, το οποίο ενημερώνει την κατάσταση των θέσεων, αναπροσαρμόζει την αξία του χαρτοφυλακίου, υπολογίζει την ανταμοιβή και παράγει την επόμενη κατάσταση Η ανταμοιβή αντικατοπτρίζει τόσο την οικονομική συνέπεια της επιλεγμένης απόφασης όσο και στοιχεία διαχείρισης κινδύνου που ενσωματώνονται στη μαθησιακή διαδικασία, επιτρέποντας στον πράκτορα να αξιολογεί τη στρατηγική του με βάση πιο ολοκληρωμένα χρηματοοικονομικά κριτήρια.

Κάθε εμπειρία μετάβασης αποθηκεύεται στη μνήμη επανάληψης εμπειριών (replay memory) υπό τη μορφή: (s, a, r, s′, done), επιτρέποντας την τυχαία επαναδειγματοληψία κατά τη διαδικασία εκπαίδευσης. Η αποθήκευση αυτή επιτρέπει την τυχαία επαναδειγματοληψία εμπειριών κατά τη διαδικασία εκπαίδευσης, μειώνοντας τη χρονική συσχέτιση διαδοχικών παρατηρήσεων και ενισχύοντας τη σταθερότητα της μάθησης.

Κατά την εκπαιδευτική διαδικασία, ο πράκτορας χρησιμοποιεί δείγματα από τη replay memory για την ενημέρωση του policy network, ενώ το target network παρέχει σταθερότερες τιμές-στόχους για τη διαδικασία βελτιστοποίησης. Η διαδικασία αυτή επαναλαμβάνεται διαρκώς σε κάθε επεισόδιο, οδηγώντας στη σταδιακή βελτίωση της στρατηγικής και στη σύγκλιση της μαθησιακής διαδικασίας.

Μετά την ολοκλήρωση συγκεκριμένων εκπαιδευτικών κύκλων, το σύστημα αξιολογείται σε νέα χρονικά παράθυρα μέσω διαδικασιών rolling window ή walk-forward validation, χωρίς περαιτέρω ενημέρωση των παραμέτρων του μοντέλου. Η προσέγγιση αυτή επιτρέπει την αντικειμενικότερη αποτίμηση της δυνατότητας γενίκευσης της στρατηγικής σε μη παρατηρημένα δεδομένα.

Συνολικά, η ροή λειτουργίας του συστήματος ακολουθεί έναν πλήρη κύκλο: εισαγωγή δεδομένων, προεπεξεργασία, διαμόρφωση κατάστασης, επιλογή ενέργειας, εφαρμογή στο

περιβάλλον, υπολογισμός ανταμοιβής, αποθήκευση εμπειρίας, εκπαίδευση και τελική αξιολόγηση. Η κυκλική αυτή διαδικασία αποτελεί τον θεμελιώδη μηχανισμό μάθησης του προτεινόμενου συστήματος.

Η συγκεκριμένη σχεδίαση επιτυγχάνει σαφή διαχωρισμό μεταξύ υποσυστημάτων εισόδου, περιβάλλοντος, πράκτορα και αξιολόγησης, ενισχύοντας τη μεθοδολογική καθαρότητα, την τεχνική διαφάνεια και την επιστημονική αξιοπιστία της υλοποίησης. Παράλληλα, διασφαλίζει ότι οι συγκριτικές αξιολογήσεις μεταξύ BASE και VWAP πραγματοποιούνται υπό σταθερές αρχιτεκτονικές συνθήκες, επιτρέποντας ακριβέστερη αποτίμηση της πραγματικής συνεισφοράς του δείκτη VWAP.

4.2 Flow chart λειτουργίας του συστήματος

Η λειτουργία του προτεινόμενου συστήματος ενισχυτικής μάθησης αποτυπώνεται αναλυτικά μέσω διαγραμμάτων ροής (flowcharts), τα οποία παρουσιάζουν τη διαδοχική εξέλιξη των διαδικασιών εκπαίδευσης και αξιολόγησης του πράκτορα. Η χρήση των flowcharts επιτρέπει την οπτική και λειτουργική αναπαράσταση της συνολικής εκτέλεσης του συστήματος, διευκολύνοντας την κατανόηση της λογικής ακολουθίας των επιμέρους σταδίων και της αλληλεπίδρασης μεταξύ των βασικών υποσυστημάτων.

Σε αντίθεση με τις προηγούμενες ενότητες, όπου παρουσιάστηκαν κυρίως η αρχιτεκτονική δομή, τα υποσυστήματα και οι θεωρητικές σχεδιαστικές επιλογές, η παρούσα ενότητα επικεντρώνεται στην επιχειρησιακή λειτουργία του συστήματος κατά τη διάρκεια της πραγματικής εκτέλεσης των πειραμάτων. Τα διαγράμματα ροής αναπαριστούν βήμα προς βήμα τη διαδικασία μέσω της οποίας το σύστημα μετατρέπει τα δεδομένα αγοράς σε αποφάσεις συναλλαγών, εκπαιδεύει τον πράκτορα και αξιολογεί τη στρατηγική του.

Συγκεκριμένα, η λειτουργική διαδικασία περιλαμβάνει τη φόρτωση και προεπεξεργασία των δεδομένων, τη δημιουργία του διανύσματος κατάστασης, τη λήψη αποφάσεων από τον πράκτορα, την εφαρμογή ενεργειών στο περιβάλλον συναλλαγών, τον υπολογισμό ανταμοιβής, την αποθήκευση εμπειριών στη replay memory, την ενημέρωση των νευρωνικών δικτύων και, τελικά, την αξιολόγηση της απόδοσης του συστήματος μέσω κατάλληλων χρηματοοικονομικών και μαθησιακών μετρικών.

Ιδιαίτερη σημασία έχει η διάκριση μεταξύ της διαδικασίας εκπαίδευσης και της διαδικασίας αξιολόγησης, καθώς οι δύο αυτές φάσεις διαφοροποιούνται ουσιαστικά ως προς τη λειτουργική τους δομή. Κατά τη διάρκεια της εκπαίδευσης, εφαρμόζεται πολιτική εξερεύνησης τύπου ε-greedy, πραγματοποιείται συνεχής ενημέρωση του policy network, χρησιμοποιείται replay memory και υλοποιείται περιοδική ενημέρωση του target network. Αντίθετα, κατά την αξιολόγηση εφαρμόζεται σταθερή greedy πολιτική, δεν πραγματοποιείται περαιτέρω εκπαίδευση και το σύστημα επικεντρώνεται αποκλειστικά στην αποτίμηση της γενικευτικής ικανότητας της ήδη εκπαιδευμένης στρατηγικής.

Η διαφοροποίηση αυτή καθιστά απαραίτητη την ανάπτυξη ξεχωριστών flowcharts για την εκπαίδευση και την αξιολόγηση, ώστε να αποτυπώνονται με σαφήνεια οι διαφορετικοί

λειτουργικοί στόχοι κάθε φάσης και να ενισχύεται η μεθοδολογική πληρότητα της παρουσίασης.

Παράλληλα, τα διαγράμματα ροής ενσωματώνουν τη λογική rolling-window ή walk-forward validation που εφαρμόζεται στα πειραματικά σενάρια της εργασίας. Σε κάθε κύκλο, το σύστημα εκπαιδεύεται σε συγκεκριμένο χρονικό παράθυρο και αξιολογείται σε επόμενη περίοδο αγοράς, προσομοιώνοντας πιο ρεαλιστικές συνθήκες διαδοχικής προσαρμογής σε μεταβαλλόμενα περιβάλλοντα αγοράς.

Σημαντικό είναι ότι οι εκδοχές BASE και VWAP διαφοροποιούνται αποκλειστικά στο στάδιο διαμόρφωσης των χαρακτηριστικών εισόδου και όχι στον βασικό μηχανισμό λειτουργίας του συστήματος. Ως αποτέλεσμα, τα flowcharts παραμένουν ουσιαστικά κοινά και για τις δύο εκδοχές, με μόνη διαφοροποίηση την παρουσία ή απουσία των VWAP χαρακτηριστικών στο state vector.

Η χρήση των flowcharts ενισχύει σημαντικά τη σαφήνεια της τεχνικής τεκμηρίωσης, διευκολύνει τη σύνδεση μεταξύ θεωρίας και πραγματικής υλοποίησης, υποστηρίζει την παρουσίαση των ψευδοκωδικών που ακολουθούν και προσφέρει μια πλήρη επιχειρησιακή αναπαράσταση του τρόπου λειτουργίας του συστήματος.

Συνολικά, τα διαγράμματα ροής αποτελούν κρίσιμο στοιχείο της τεχνικής και ακαδημαϊκής τεκμηρίωσης της εργασίας, καθώς αποτυπώνουν με σαφή, οργανωμένο και μεθοδολογικά συνεπή τρόπο τη διαδικασία μέσω της οποίας ο πράκτορας συλλέγει πληροφορία, λαμβάνει αποφάσεις, εκπαιδεύεται και αξιολογείται σε συνθήκες αυτόματης ενδοημερήσιας διαπραγμάτευσης.

4.2.1 Διάγραμμα ροής της διαδικασίας εκπαίδευσης

Το διάγραμμα ροής της διαδικασίας εκπαίδευσης αποτυπώνει τη λειτουργική ακολουθία μέσω της οποίας ο πράκτορας Dueling DQN αλληλεπιδρά με το περιβάλλον συναλλαγών και προσαρμόζει σταδιακά τη στρατηγική λήψης αποφάσεων. Η διαδικασία ξεκινά με τη φόρτωση των ιστορικών δεδομένων αγοράς, τα οποία περιλαμβάνουν πληροφορίες τιμών Open, High, Low, Close και Volume, καθώς και τα τεχνικά χαρακτηριστικά που αξιοποιούνται για τη δημιουργία της αναπαράστασης κατάστασης.

Στο στάδιο της προεπεξεργασίας, τα δεδομένα καθαρίζονται, κανονικοποιούνται και μετασχηματίζονται σε διανύσματα κατάστασης κατάλληλα για είσοδο στο νευρωνικό δίκτυο. Στην εκδοχή VWAP, το διάνυσμα κατάστασης (state vector) εμπλουτίζεται με πρόσθετα χαρακτηριστικά που σχετίζονται με τον δείκτη VWAP, ενώ στην εκδοχή BASE η αρχιτεκτονική παραμένει ίδια χωρίς την ενσωμάτωση της συγκεκριμένης πληροφορίας. Με τον τρόπο αυτό διασφαλίζεται ότι οι συγκρίσεις μεταξύ των δύο σεναρίων πραγματοποιούνται υπό κοινό εκπαιδευτικό πλαίσιο.

Σε επίπεδο πρακτικής υλοποίησης, το διάνυσμα κατάστασης του πράκτορα σε κάθε χρονικό βήμα διαμορφώνεται ως συνδυασμός κανονικοποιημένων χαρακτηριστικών αγοράς, τεχνικών δεικτών και μεταβλητών που αποτυπώνουν τη θέση του πράκτορα.

Συγκεκριμένα, στις δύο πειραματικές εκδοχές του συστήματος ορίζεται ως εξής:

$$state_{BASE(t)} = [x(t), pos(t)] \qquad (4.2)$$

$$state_{VWAP(t)} = [x(t), vwap_{features(t)}, pos(t)] \qquad (4.3)$$

όπου:

x(t): διάνυσμα κανονικοποιημένων χαρακτηριστικών αγοράς και τεχνικών δεικτών στη χρονική στιγμή t,

$vwap_{features(t)}$: σύνολο χαρακτηριστικών που σχετίζονται με τον δείκτη VWAP, όπως η τιμή του VWAP ή/και η σχετική απόκλιση της τρέχουσας τιμής από αυτόν,

pos(t): σύνολο μεταβλητών που περιγράφουν την τρέχουσα κατάσταση του πράκτορα και του χαρτοφυλακίου του, όπως η ύπαρξη ή μη ανοικτής θέσης, ο τύπος της θέσης (long ή short), καθώς και λοιπές πληροφορίες σχετικές με τη διαχείριση της επενδυτικής θέσης.

Η συγκεκριμένη αναπαράσταση επιτρέπει τη συνδυαστική αξιοποίηση πληροφορίας αγοράς, τεχνικής ανάλυσης και κατάστασης χαρτοφυλακίου, διαμορφώνοντας ένα συνεκτικό πληροφοριακό πλαίσιο για τη διαδικασία λήψης αποφάσεων του πράκτορα.

Σε κάθε χρονικό βήμα του επεισοδίου, το περιβάλλον παρέχει την τρέχουσα κατάσταση στον πράκτορα, ο οποίος επιλέγει ενέργεια μέσω πολιτικής ε-greedy. Η στρατηγική αυτή επιτρέπει ισορροπία μεταξύ εξερεύνησης νέων ενεργειών και εκμετάλλευσης ήδη μαθημένων πολιτικών.

Η επιλεγμένη ενέργεια εφαρμόζεται στο περιβάλλον συναλλαγών, το οποίο ενημερώνει τις ανοικτές θέσεις, την αξία του χαρτοφυλακίου και τη νέα κατάσταση της αγοράς. Στη συνέχεια υπολογίζεται η ανταμοιβή, η οποία βασίζεται στη μεταβολή της αξίας του χαρτοφυλακίου, αποτυπώνοντας την οικονομική συνέπεια της επιλεγμένης στρατηγικής.

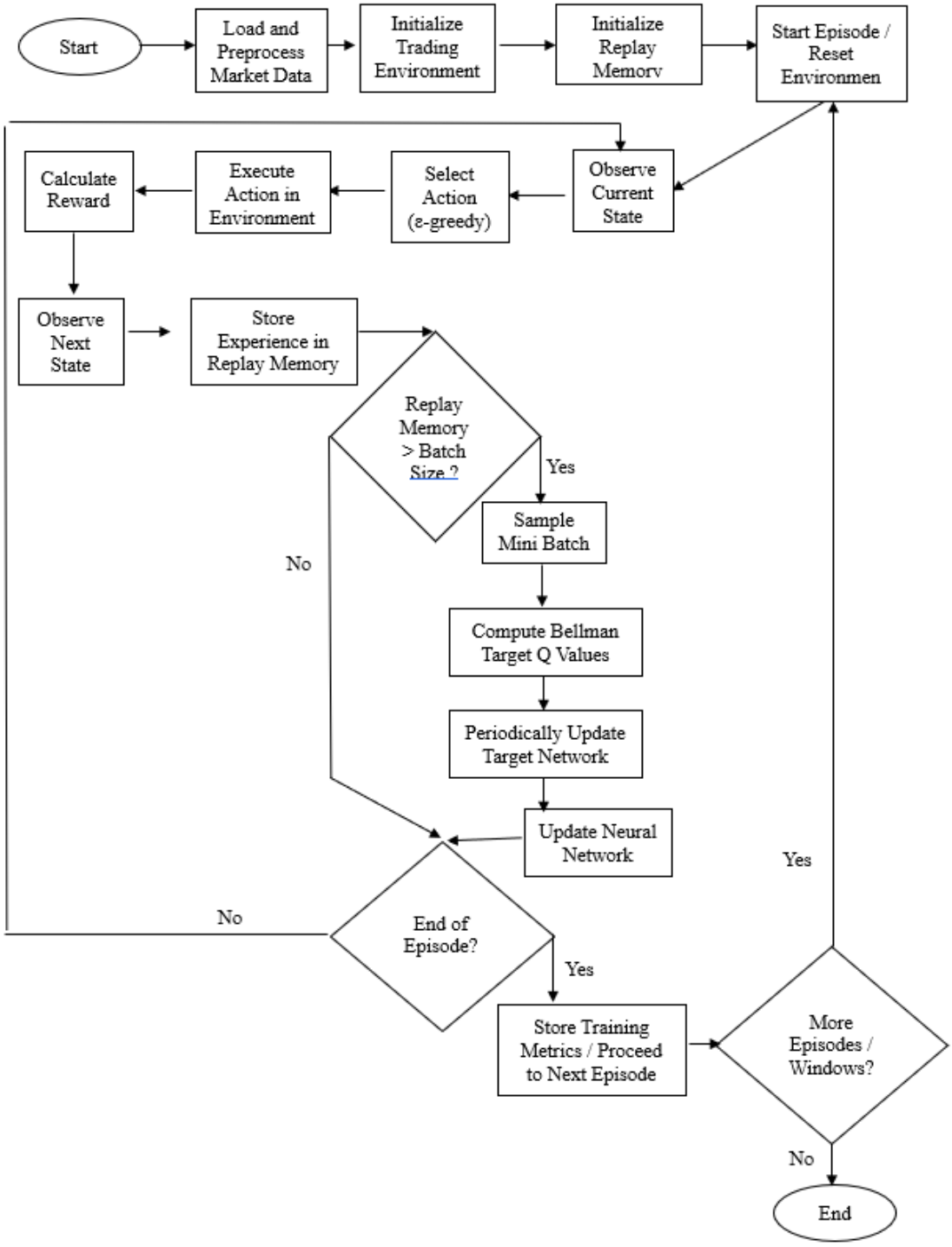
Η εμπειρία κάθε χρονικού βήματος αποθηκεύεται στη μνήμη επανάληψης εμπειριών (replay memory) με τη μορφή (s, a, r, s′, done), όπου s είναι η τρέχουσα κατάσταση, a η επιλεγμένη ενέργεια, r η ανταμοιβή, s′ η επόμενη κατάσταση και done η ένδειξη ολοκλήρωσης του επεισοδίου.

Κατά την εκπαιδευτική διαδικασία, τυχαία δείγματα εμπειριών αντλούνται από τη replay memory και χρησιμοποιούνται για την ενημέρωση του policy network. Η διαδικασία αυτή μειώνει τη χρονική συσχέτιση μεταξύ διαδοχικών παρατηρήσεων και ενισχύει τη σταθερότητα της μάθησης.

Παράλληλα, το target network ενημερώνεται περιοδικά ώστε να παρέχει σταθερότερες τιμές-στόχους κατά τον υπολογισμό της συνάρτησης απώλειας. Η επαναληπτική αυτή διαδικασία συνεχίζεται έως την ολοκλήρωση κάθε επεισοδίου και επαναλαμβάνεται σε διαδοχικά χρονικά παράθυρα σύμφωνα με τη λογική rolling-window ή walk-forward.

```
1. Αρχικοποίηση Policy Network Q_policy(s,a;θ)
2. Αρχικοποίηση Target Network Q_target(s,a;θ⁻)
3. Αντιγραφή παραμέτρων: θ⁻ ← θ
4. Αρχικοποίηση Replay Memory D
5. Ορισμός αρχικής τιμής ε
6. Για κάθε rolling-window / επεισόδιο:
   6.1 Φόρτωση δεδομένων αγοράς
   6.2 Προεπεξεργασία και δημιουργία state vectors
   6.3 Αρχικοποίηση Trading Environment
   6.4 Λήψη αρχικής κατάστασης s
   6.5 Όσο done = False:
       6.5.1 Με πιθανότητα ε:
             Επιλογή τυχαίας ενέργειας a
       6.5.2 Αλλιώς:
             a = argmax_a Q_policy(s,a;θ)
       6.5.3 Εκτέλεση ενέργειας a στο περιβάλλον
       6.5.4 Λήψη νέας κατάστασης s', ανταμοιβής r και ένδειξης done
       6.5.5 Αποθήκευση εμπειρίας (s, a, r, s', done) → D
       6.5.6 Αν |D| > batch_size:
             Δειγματοληψία mini-batch από D
             Για κάθε εμπειρία:
                Αν done = True:
                   y = r
                Αλλιώς:
                   y = r + γ max_a' Q_target(s',a';θ⁻)
                Υπολογισμός L(θ) = (Q_policy(s,a;θ) - y)^2
             Ενημέρωση θ μέσω optimizer
       6.5.7 Αν συμπληρωθεί target_update_interval:
             θ⁻ ← θ
       6.5.8 s ← s'
   6.6 Μείωση ε
7. Αποθήκευση τελικού μοντέλου
```

Ψευδοκώδικας 1 Διαδικασία εκπαίδευσης πράκτορα Dueling DQN



Σχήμα 2 Διάγραμμα ροής της διαδικασίας εκπαίδευσης του πράκτορα Dueling Deep Q-Network σε περιβάλλον χρηματοοικονομικών συναλλαγών (BASE και VWAP εκδοχές)

4.2.2 Διάγραμμα ροής της διαδικασίας αξιολόγησης

Το διάγραμμα ροής της διαδικασίας αξιολόγησης παρουσιάζει τον τρόπο με τον οποίο εφαρμόζεται ο ήδη εκπαιδευμένος πράκτορας Dueling DQN σε νέα δεδομένα αγοράς, χωρίς περαιτέρω ενημέρωση των παραμέτρων του νευρωνικού δικτύου. Σκοπός της διαδικασίας είναι η αντικειμενική αποτίμηση της ικανότητας γενίκευσης της στρατηγικής που αναπτύχθηκε κατά τη φάση της εκπαίδευσης.

Η διαδικασία ξεκινά με τη φόρτωση του εκπαιδευμένου policy network και των δεδομένων αξιολόγησης. Στη συνέχεια, τα δεδομένα υποβάλλονται στις ίδιες διαδικασίες προεπεξεργασίας και κανονικοποίησης που εφαρμόστηκαν κατά την εκπαίδευση, χρησιμοποιώντας τις ίδιες παραμέτρους μετασχηματισμού, ώστε να διασφαλίζεται μεθοδολογική συνέπεια και να αποφεύγεται διαρροή πληροφορίας.

Ακολούθως, αρχικοποιείται το περιβάλλον αξιολόγησης και δημιουργείται το αρχικό διάνυσμα κατάστασης state(t), το οποίο αποτελεί είσοδο στο policy network. Ανάλογα με το εξεταζόμενο σενάριο, το state vector περιλαμβάνει ή όχι τα χαρακτηριστικά του δείκτη VWAP, χωρίς να τροποποιείται η βασική αρχιτεκτονική του συστήματος.

Κατά τη φάση αξιολόγησης, η επιλογή ενεργειών πραγματοποιείται αποκλειστικά μέσω greedy policy. Ο πράκτορας επιλέγει την ενέργεια με τη μέγιστη εκτιμώμενη τιμή Q(s,a), χωρίς χρήση μηχανισμού εξερεύνησης. Συνεπώς:

• δεν εφαρμόζεται ε-greedy στρατηγική

• δεν πραγματοποιείται αποθήκευση εμπειριών

• δεν χρησιμοποιείται replay memory για εκπαίδευση

• δεν πραγματοποιείται ενημέρωση των παραμέτρων του δικτύου

• δεν ενημερώνεται το target network

Σε κάθε χρονικό βήμα, η επιλεγμένη ενέργεια εφαρμόζεται στο περιβάλλον συναλλαγών, το οποίο ενημερώνει την κατάσταση των θέσεων, υπολογίζει τη νέα αξία του χαρτοφυλακίου και επιστρέφει την επόμενη κατάσταση. Παράλληλα, καταγράφονται οι πραγματοποιούμενες συναλλαγές και η εξέλιξη της αξίας του χαρτοφυλακίου.

Κατά τη φάση αξιολόγησης, η ανταμοιβή εξακολουθεί να υπολογίζεται με τον ίδιο τρόπο όπως κατά την εκπαίδευση, ώστε να καταγράφεται η συμπεριφορά της στρατηγικής, δεν χρησιμοποιείται όμως για περαιτέρω ενημέρωση των παραμέτρων του μοντέλου.

Η διαδικασία επαναλαμβάνεται έως την ολοκλήρωση του επεισοδίου (ή του χρονικού παραθύρου αξιολόγησης), οπότε και υπολογίζεται η τελική αξία του χαρτοφυλακίου.

Με την ολοκλήρωση της αξιολόγησης παράγονται οι βασικές μετρικές απόδοσης, οι οποίες περιλαμβάνουν:

• εξέλιξη αξίας χαρτοφυλακίου (portfolio curve)

• συνολικό Profit and Loss (PnL)

• σωρευτική απόδοση

• αριθμό και συχνότητα συναλλαγών

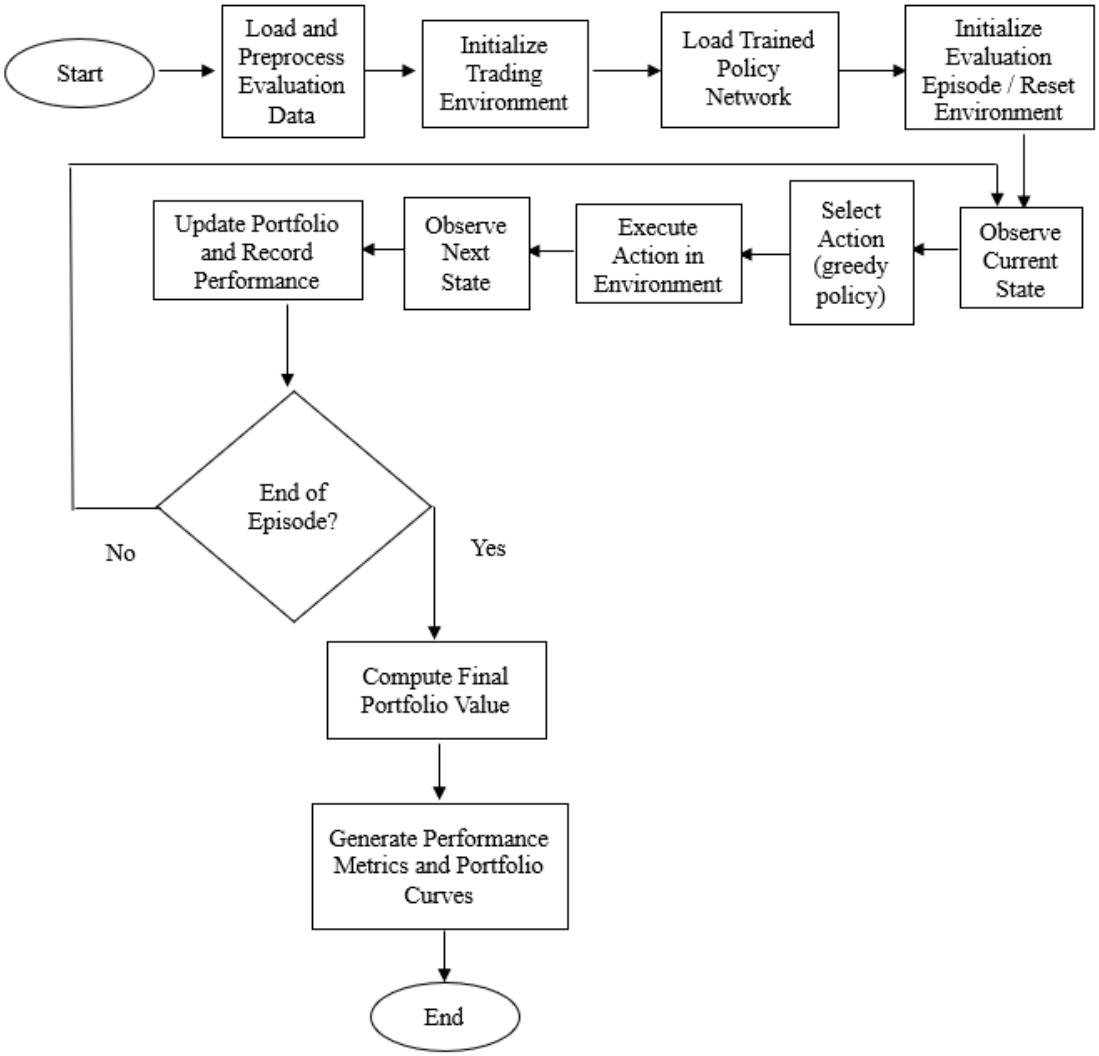
• ιστορικό συναλλαγών

• συγκριτικές μετρικές μεταξύ των εκδοχών BASE και VWAP

Η διαδικασία αξιολόγησης οργανώνεται με βάση λογική rolling-window ή walk-forward validation, όπου το εκπαιδευμένο μοντέλο εφαρμόζεται σε διαδοχικά μη επικαλυπτόμενα χρονικά παράθυρα. Η προσέγγιση αυτή προσομοιώνει ρεαλιστικές συνθήκες αγοράς και επιτρέπει αξιόπιστη εκτίμηση της ικανότητας γενίκευσης του πράκτορα.

```
1. Φόρτωση εκπαιδευμένου Policy Network Q_policy(s,a;θ)
2. Φόρτωση δεδομένων αξιολόγησης
3. Προεπεξεργασία δεδομένων και δημιουργία state vectors
4. Αρχικοποίηση Trading Environment αξιολόγησης
5. Λήψη αρχικής κατάστασης s
6. Αρχικοποίηση λίστας συναλλαγών και ιστορικού χαρτοφυλακίου
7. Όσο done = False:
   7.1 Επιλογή ενέργειας με greedy policy:
       a = argmax_a Q_policy(s,a;θ)
   7.2 Εκτέλεση ενέργειας a στο περιβάλλον
   7.3 Λήψη νέας κατάστασης s', ανταμοιβής r και ένδειξης done
   7.4 Ενημέρωση αξίας χαρτοφυλακίου
   7.5 Καταγραφή συναλλαγής, εφόσον πραγματοποιήθηκε
   7.6 Καταγραφή τρέχουσας αξίας χαρτοφυλακίου
   7.7 s ← s'
8. Υπολογισμός τελικών μετρικών απόδοσης
9. Παραγωγή portfolio curve και ιστορικού συναλλαγών
10. Σύγκριση αποτελεσμάτων BASE και VWAP
11. Αποθήκευση αποτελεσμάτων και διαγραμμάτων
```

Ψευδοκώδικας 2 Διαδικασία αξιολόγησης πράκτορα Dueling DQN



Σχήμα 3 Διάγραμμα ροής της διαδικασίας αξιολόγησης του πράκτορα Dueling Deep Q-Network σε περιβάλλον χρηματοοικονομικών συναλλαγών (BASE και VWAP εκδοχές)

Συνολικά, το διάγραμμα ροής της αξιολόγησης αποτυπώνει τη μετάβαση από τη μαθησιακή διαδικασία στην πρακτική εφαρμογή της στρατηγικής. Η σαφής διάκριση μεταξύ εκπαίδευσης και αξιολόγησης ενισχύει τη μεθοδολογική εγκυρότητα της εργασίας και επιτρέπει αντικειμενική σύγκριση των πειραματικών εκδοχών.

4.2.3 Περιγραφή των βασικών σταδίων λειτουργίας

Η λειτουργία του προτεινόμενου συστήματος ενισχυτικής μάθησης βασίζεται σε μια ολοκληρωμένη αλληλουχία διαδοχικών σταδίων, τα οποία επιτρέπουν τη συστηματική εκπαίδευση, αξιολόγηση και συγκριτική αποτίμηση του πράκτορα Dueling DQN σε διαφορετικά πειραματικά σενάρια. Τα στάδια αυτά συνθέτουν έναν πλήρη λειτουργικό κύκλο, όπου τα ιστορικά δεδομένα αγοράς μετατρέπονται σε πληροφοριακές αναπαραστάσεις, οι οποίες οδηγούν σε αποφάσεις συναλλαγών, ενώ τα αποτελέσματα των αποφάσεων αυτών αξιοποιούνται για τη συνεχή προσαρμογή και βελτίωση της στρατηγικής.

Η διαδικασία ξεκινά με την εισαγωγή ιστορικών χρηματοοικονομικών δεδομένων αγοράς, τα οποία περιλαμβάνουν τιμές Open, High, Low, Close και Volume, καθώς και τεχνικούς δείκτες και, όπου εφαρμόζεται, χαρακτηριστικά που σχετίζονται με τον δείκτη VWAP. Τα δεδομένα αυτά αποτελούν τη βασική εξωτερική πληροφοριακή πηγή του συστήματος και τροφοδοτούν όλες τις επόμενες διαδικασίες.

Στη συνέχεια, εφαρμόζονται διαδικασίες προεπεξεργασίας και εξαγωγής χαρακτηριστικών, οι οποίες περιλαμβάνουν καθαρισμό, κανονικοποίηση, υπολογισμό τεχνικών χαρακτηριστικών και μετασχηματισμό των πρωτογενών δεδομένων σε μορφή κατάλληλη για το νευρωνικό δίκτυο. Το στάδιο αυτό είναι κρίσιμο για τη διασφάλιση αριθμητικής σταθερότητας και πληροφοριακής ποιότητας.

Για κάθε χρονικό βήμα, τα διαθέσιμα χαρακτηριστικά οργανώνονται σε ένα διάνυσμα κατάστασης (state vector), το οποίο αποτελεί τη συνοπτική πληροφοριακή αναπαράσταση της αγοράς και της τρέχουσας θέσης του πράκτορα. Το state vector περιλαμβάνει στοιχεία σχετικά με τη δυναμική της αγοράς, τεχνικά χαρακτηριστικά, πληροφορία χαρτοφυλακίου και, όπου εφαρμόζεται, πρόσθετη πληροφορία VWAP.

Ο πράκτορας επεξεργάζεται το state vector μέσω του policy network και επιλέγει μία από τις διαθέσιμες ενέργειες αγοράς, πώλησης ή διακράτησης. Κατά την εκπαιδευτική διαδικασία χρησιμοποιείται στρατηγική ε-greedy για την ισορροπία μεταξύ εξερεύνησης και εκμετάλλευσης, ενώ κατά τη φάση αξιολόγησης εφαρμόζεται αποκλειστικά greedy policy.

Η επιλεγμένη ενέργεια εφαρμόζεται στο περιβάλλον συναλλαγών, το οποίο ενημερώνει τις θέσεις του χαρτοφυλακίου, καταγράφει τις συναλλαγές, υπολογίζει τη νέα αξία του χαρτοφυλακίου και δημιουργεί την επόμενη κατάσταση του συστήματος. Το περιβάλλον αυτό

προσομοιώνει τη ρεαλιστική δυναμική της αγοράς και λειτουργεί ως ο βασικός μηχανισμός αλληλεπίδρασης agent–environment.

Μετά την εφαρμογή της ενέργειας, υπολογίζεται η ανταμοιβή του πράκτορα, η οποία συνδυάζει στοιχεία απόδοσης και διαχείρισης κινδύνου. Η ανταμοιβή διαμορφώνεται με βάση το οικονομικό αποτέλεσμα των συναλλαγών, τη μεταβολή της αξίας του χαρτοφυλακίου και μηχανισμούς προσαρμογής κινδύνου, όπως αναλύθηκε στην Ενότητα 3.3. Κατά τη φάση εκπαίδευσης, η ανταμοιβή αυτή χρησιμοποιείται ως μαθησιακό σήμα για την ενημέρωση του μοντέλου, ενώ κατά τη φάση αξιολόγησης χρησιμοποιείται αποκλειστικά για σκοπούς καταγραφής και ανάλυσης.

Κατά τη φάση εκπαίδευσης, κάθε εμπειρία αποθηκεύεται στη replay memory με τη μορφή (s, a, r, s′, done), επιτρέποντας την επαναδειγματοληψία και τη σταθεροποίηση της μαθησιακής διαδικασίας.

Η χρήση της μνήμης εμπειριών επιτρέπει τυχαία δειγματοληψία κατά την εκπαίδευση, περιορίζοντας τη χρονική εξάρτηση των παρατηρήσεων και ενισχύοντας τη σταθερότητα της διαδικασίας μάθησης.

Το policy network ενημερώνεται μέσω επαναλαμβανόμενης διαδικασίας βελτιστοποίησης, η οποία περιλαμβάνει batch sampling, υπολογισμό συνάρτησης απώλειας, gradient descent και περιοδικό συγχρονισμό με το target network. Η διαδικασία αυτή οδηγεί στη σταδιακή βελτίωση της στρατηγικής του πράκτορα.

Με την ολοκλήρωση κάθε χρονικού παραθύρου ή επεισοδίου, το σύστημα μεταβαίνει στην επόμενη περίοδο δεδομένων σύμφωνα με τη διάρθρωση του εκάστοτε πειραματικού σεναρίου. Στο εκτεταμένο σενάριο αξιολόγησης εφαρμόζεται πλαίσιο rolling-window και walk-forward validation, επιτρέποντας τη διαδοχική αξιολόγηση της στρατηγικής σε νέα δεδομένα αγοράς. Η προσέγγιση αυτή προσομοιώνει πιο ρεαλιστικά πραγματικές συνθήκες αλγοριθμικής διαπραγμάτευσης.

Στην τελική φάση αξιολόγησης, το εκπαιδευμένο μοντέλο εφαρμόζεται σε νέα δεδομένα αγοράς και παράγει αποτελέσματα όπως:

• εξέλιξη χαρτοφυλακίου

• ιστορικό συναλλαγών

• μετρικές απόδοσης

• συγκριτικές αναλύσεις BASE και VWAP

Συνολικά, η λειτουργία του συστήματος μπορεί να συνοψιστεί ως εξής:

Δεδομένα αγοράς

→ Προεπεξεργασία

→ Δημιουργία κατάστασης

→ Πράκτορας Dueling DQN

→ Επιλογή ενέργειας

→ Περιβάλλον συναλλαγών

→ Υπολογισμός ανταμοιβής

→ Replay memory

→ Εκπαίδευση

→ Αξιολόγηση

→ Αποτελέσματα

Η διαδικασία αυτή είναι επαναληπτική, δυναμική και πλήρως προσαρμοσμένη στις απαιτήσεις ενός προηγμένου συστήματος αυτόματης χρηματοοικονομικής διαπραγμάτευσης.

Συμπερασματικά, τα βασικά στάδια λειτουργίας του συστήματος αποτυπώνουν έναν πλήρη μηχανισμό βαθιάς ενισχυτικής μάθησης, όπου η συνεχής αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος οδηγεί στη σταδιακή ανάπτυξη, προσαρμογή και αξιολόγηση αποδοτικών στρατηγικών συναλλαγών. Η προσέγγιση αυτή διασφαλίζει τεχνική συνοχή, μεθοδολογική εγκυρότητα και ισχυρή σύνδεση μεταξύ θεωρητικού πλαισίου και πρακτικής υλοποίησης.

4.3Υλοποίηση πράκτορα Dueling Deep Q-Network

4.3.1 Θεμελιώδεις μαθηματικές σχέσεις

Η υλοποίηση του πράκτορα Dueling Deep Q-Network βασίζεται στις θεμελιώδεις αρχές της βαθιάς ενισχυτικής μάθησης και ειδικότερα στο πλαίσιο του Q-learning, όπου ο στόχος είναι η εκμάθηση μιας πολιτικής επιλογής ενεργειών που μεγιστοποιεί τη σωρευτική μελλοντική ανταμοιβή κατά την αλληλεπίδραση του πράκτορα με το περιβάλλον συναλλαγών.

Στο πλαίσιο αυτό, ο πράκτορας επιδιώκει να προσεγγίσει τη συνάρτηση δράσης-αξίας Q(s,a), η οποία εκφράζει την αναμενόμενη συνολική ανταμοιβή όταν βρίσκεται σε κατάσταση s και επιλέγει ενέργεια a.

Στην αρχιτεκτονική Dueling DQN, η συνάρτηση αυτή αποσυντίθεται σε δύο επιμέρους συνιστώσες: την αξία της κατάστασης V(s) και το πλεονέκτημα της ενέργειας A(s,a). Η πρακτική μαθηματική διατύπωση λαμβάνει τη μορφή:

$$Q(s,a) = V(s) + A(s,a) - mean_{\alpha\ A(s,a)} \qquad (4.4)$$

όπου:

Q(s,a): συνολική εκτίμηση αξίας επιλογής ενέργειας,

V(s): εκτίμηση της συνολικής αξίας της κατάστασης,

A(s,a): σχετικό πλεονέκτημα της συγκεκριμένης ενέργειας,

$mean_{\alpha\ A(s,a)}$: μέσο πλεονέκτημα όλων των διαθέσιμων ενεργειών στην κατάσταση s.

Η αφαίρεση του μέσου όρου του πλεονεκτήματος διασφαλίζει σταθερότερη αριθμητική συμπεριφορά και αποτρέπει προβλήματα μη μοναδικότητας κατά την εκπαίδευση.

Κατά τη διαδικασία μάθησης, η τιμή-στόχος υπολογίζεται μέσω της εξίσωσης Bellman,

y = r + γ max_a' Q_target(s', a'; θ').

η οποία επιτρέπει στον πράκτορα να ενσωματώνει τόσο την άμεση ανταμοιβή όσο και τις εκτιμώμενες μελλοντικές συνέπειες των αποφάσεών του.

Η βελτιστοποίηση του policy network πραγματοποιείται μέσω ελαχιστοποίησης της συνάρτηση

$$L(\theta) = (\text{Q\_policy}(s, a; \theta)\ - y)\text{^}2 \qquad (4.5)$$

όπου:

L(θ): συνάρτηση απώλειας,

θ: παράμετροι του policy network,

Q_policy(s,a;θ): τρέχουσα εκτίμηση της συνάρτησης Q,

y: τιμή-στόχος που προκύπτει από την εξίσωση Bellman.

Η ελαχιστοποίηση της συνάρτησης αυτής οδηγεί στη σταδιακή προσαρμογή των βαρών του δικτύου και στη βελτίωση της στρατηγικής.

Η συνάρτηση ανταμοιβής του πράκτορα δεν βασίζεται αποκλειστικά στη μεταβολή της αξίας του χαρτοφυλακίου, αλλά συνδυάζει στοιχεία απόδοσης και διαχείρισης κινδύνου. Η τελική ανταμοιβή προκύπτει από τρεις βασικές συνιστώσες: το οικονομικό αποτέλεσμα των συναλλαγών, την προσαρμογή βάσει του δείκτη Sortino και την επιβολή ποινής λόγω drawdown.

Αρχικά υπολογίζεται η βασική ανταμοιβή:

$$r_{base(t)} = r_{pnl(t)} - TC_t \qquad (4.6)$$

όπου:

$r_{base(t)}$: βασική ανταμοιβή στο χρονικό βήμα t,

$r_{pnl(t)}$: συνιστώσα κέρδους ή ζημίας (Profit and Loss) που προκύπτει από τη μεταβολή της αξίας του χαρτοφυλακίου,

TC_t: κόστος συναλλαγών που επιβάλλεται κατά την εκτέλεση ενεργειών αγοράς ή πώλησης.

Στη συνέχεια υπολογίζεται η προσαρμογή κινδύνου μέσω του δείκτη Sortino:

$$r_{sortino(t)} = \beta\ Sortino_t \qquad (4.7)$$

όπου:

$r_{sortino(t)}$: συνιστώσα ανταμοιβής που σχετίζεται με την απόδοση προσαρμοσμένη ως προς τον καθοδικό κίνδυνο,

β: συντελεστής βαρύτητας της συνιστώσας Sortino,

$Sortino_t$: τιμή του δείκτη Sortino στο χρονικό βήμα t.

Παράλληλα, υπολογίζεται ποινή λόγω υποχώρησης του χαρτοφυλακίου (drawdown):

$$r_{dd(t)} = \gamma\ DD_t \qquad (4.8)$$

όπου:

$r_{dd(t)}$: συνιστώσα ποινής λόγω drawdown,

γ: συντελεστής βαρύτητας της ποινής κινδύνου,

DD_t: ποσοστό drawdown του χαρτοφυλακίου στο χρονικό βήμα t.

Η τελική ανταμοιβή που χρησιμοποιείται από τον πράκτορα προκύπτει ως:

$$r_t = r_{base(t)} + r_{sortino(t)} - r_{dd(t)} \qquad (4.9)$$

όπου:

r_t: συνολική ανταμοιβή στο χρονικό βήμα t.

Η διατύπωση αυτή επιτρέπει στον πράκτορα να επιδιώκει όχι μόνο τη μεγιστοποίηση της κερδοφορίας αλλά και τον περιορισμό του κινδύνου. Με τον τρόπο αυτό, η μαθησιακή διαδικασία λαμβάνει υπόψη τόσο την απόδοση των συναλλαγών όσο και τη σταθερότητα της στρατηγικής, ενσωματώνοντας στοιχεία risk-aware reinforcement learning στη διαδικασία εκπαίδευσης.

Η επιλογή ενεργειών κατά την εκπαίδευση βασίζεται στην πολιτική ε-greedy, σύμφωνα με την οποία:

- με πιθανότητα ε επιλέγεται τυχαία ενέργεια (exploration),
- με πιθανότητα 1−ε επιλέγεται η ενέργεια με τη μέγιστη εκτιμώμενη Q-value (exploitation).

Η στρατηγική αυτή επιτρέπει ισορροπία μεταξύ εξερεύνησης του χώρου ενεργειών και αξιοποίησης της αποκτηθείσας γνώσης.

Η συνολική σταθερότητα της εκπαίδευσης ενισχύεται μέσω της χρήσης δύο νευρωνικών δικτύων:

Policy Network:

- ενημερώνεται συνεχώς,
- χρησιμοποιείται για την επιλογή ενεργειών.

Target Network:

- ενημερώνεται περιοδικά,
- χρησιμοποιείται για τον υπολογισμό των τιμών-στόχων.

Ο διαχωρισμός αυτός μειώνει σημαντικά τις αστάθειες της εκπαίδευσης και περιορίζει το φαινόμενο υπερεκτίμησης των Q-values.

Συνολικά, οι παραπάνω μαθηματικές σχέσεις αποτελούν το θεωρητικό και υπολογιστικό υπόβαθρο της υλοποίησης του πράκτορα Dueling DQN, συνδέοντας άμεσα τη θεωρία της ενισχυτικής μάθησης με την πρακτική υλοποίηση του συστήματος. Η ενσωμάτωσή τους ενισχύει τη θεωρητική συνέπεια, την ακαδημαϊκή πληρότητα και τη σαφή αντιστοίχιση μεταξύ μεθοδολογίας, υλοποίησης και πειραματικής αξιολόγησης.

4.3.2 Αρχιτεκτονική του Dueling Deep Q-Network

Η αρχιτεκτονική του πράκτορα Dueling Deep Q-Network αποτελεί τον βασικό υπολογιστικό πυρήνα του προτεινόμενου συστήματος ενισχυτικής μάθησης, καθώς είναι υπεύθυνη για την εκτίμηση της συνάρτησης δράσης-αξίας Q(s,a) και, κατά συνέπεια, για τη λήψη αποφάσεων αγοράς, πώλησης ή διακράτησης στο περιβάλλον συναλλαγών. Στόχος της είναι η μεγιστοποίηση της συνολικής μελλοντικής ανταμοιβής μέσω της εκμάθησης αποδοτικών στρατηγικών επιλογής ενεργειών.

Σε αντίθεση με ένα κλασικό Deep Q-Network, όπου το νευρωνικό δίκτυο υπολογίζει άμεσα τις τιμές Q(s,a), η αρχιτεκτονική Dueling DQN διαχωρίζει την εκτίμηση της συνάρτησης δράσης-αξίας σε δύο επιμέρους συνιστώσες: την αξία της κατάστασης V(s) και το πλεονέκτημα κάθε ενέργειας A(s,a). Η προσέγγιση αυτή επιτρέπει στο δίκτυο να εκτιμά ανεξάρτητα πόσο ευνοϊκή είναι συνολικά μια κατάσταση της αγοράς και ποια ενέργεια υπερέχει σχετικώς έναντι των εναλλακτικών επιλογών.

Η συνολική συνάρτηση δράσης-αξίας προκύπτει από τον συνδυασμό των δύο αυτών συνιστωσών, Q(s,a) = V(s) + A(s,a) − mean_a A(s,a), όπου mean_a A(s,a) είναι ο μέσος όρος του πλεονεκτήματος ως προς όλες τις διαθέσιμες ενέργειες στην κατάσταση s.

Ο διαχωρισμός αυτός βελτιώνει τη σταθερότητα της μάθησης, μειώνει την αριθμητική ασάφεια και επιτρέπει αποδοτικότερη εκτίμηση ακόμη και σε περιβάλλοντα όπου οι διαφορές μεταξύ ενεργειών είναι περιορισμένες.

Η είσοδος του πράκτορα αποτελείται από το διάνυσμα κατάστασης (state vector), το οποίο περιλαμβάνει χαρακτηριστικά τιμών αγοράς, τεχνικούς δείκτες, πληροφορίες σχετικές με την κατάσταση του χαρτοφυλακίου και, στην εκδοχή VWAP, πρόσθετα χαρακτηριστικά που σχετίζονται με τον δείκτη VWAP.

Αρχικά, το διάνυσμα εισόδου επεξεργάζεται μέσω κοινών κρυφών πλήρως συνδεδεμένων επιπέδων (shared fully connected hidden layers), τα οποία επιτελούν λειτουργίες εξαγωγής χαρακτηριστικών, μη γραμμικού μετασχηματισμού και συμπίεσης πληροφορίας.

Η πρώτη ροή, γνωστή ως Value Stream, εκτιμά τη συνολική αξία της κατάστασης V(s), ανεξάρτητα από τη συγκεκριμένη ενέργεια.

Η δεύτερη ροή, γνωστή ως Advantage Stream, εκτιμά το πλεονέκτημα κάθε διαθέσιμης ενέργειας A(s,a).

Οι δύο ροές συνδυάζονται στο τελικό επίπεδο του δικτύου, όπου υπολογίζονται οι τιμές Q(s,a) για όλες τις διαθέσιμες ενέργειες. Από τις τιμές αυτές προκύπτει η τελική επιλογή ενέργειας:

$$a = argmax\ Q(s,a) \qquad (4.10)$$

Κατά τη φάση εκπαίδευσης εφαρμόζεται πολιτική ε-greedy, ενώ κατά τη φάση αξιολόγησης εφαρμόζεται καθαρά greedy πολιτική.

Η υλοποίηση περιλαμβάνει δύο πανομοιότυπα νευρωνικά δίκτυα:

• Policy Network, το οποίο ενημερώνεται συνεχώς και χρησιμοποιείται για επιλογή ενεργειών,

• Target Network, το οποίο ενημερώνεται περιοδικά και χρησιμοποιείται για τον υπολογισμό σταθερότερων τιμών-στόχων.

Η χρήση των δύο αυτών δικτύων ενισχύει σημαντικά τη σταθερότητα της εκπαίδευσης και περιορίζει φαινόμενα υπερεκτίμησης Q-values.

Η συγκεκριμένη αρχιτεκτονική προσφέρει σημαντικά πλεονεκτήματα για χρηματοοικονομικές εφαρμογές, όπως:

• βελτιωμένη αξιολόγηση καταστάσεων αγοράς,

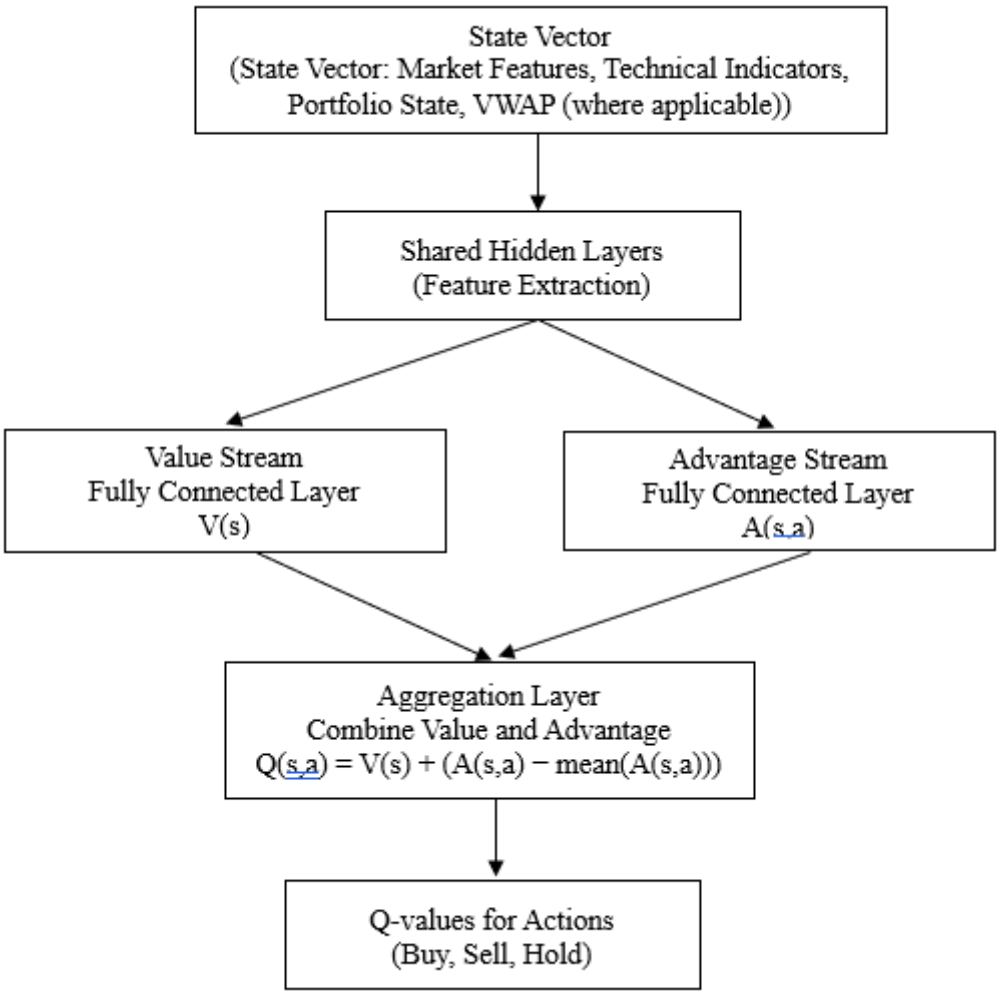
• σταθερότερη μαθησιακή διαδικασία,

• αποδοτικότερη σύγκλιση,

• αυξημένη γενικευσιμότητα,

• καταλληλότητα για σύνθετα περιβάλλοντα αλγοριθμικής διαπραγμάτευσης.

Στην υλοποίηση της παρούσας εργασίας, η αρχιτεκτονική αυτή εφαρμόζεται άμεσα μέσω των επιμέρους υπολογιστικών υπομονάδων του νευρωνικού δικτύου (neural network modules) και των μηχανισμών εκπαίδευσης. Παραμένει κοινή τόσο για το BASE όσο και για το VWAP σενάριο, επιτρέποντας αυστηρή συγκριτική αξιολόγηση της επίδρασης του VWAP χωρίς τροποποίηση του βασικού μηχανισμού μάθησης.



Σχήμα 4 Αρχιτεκτονική του πράκτορα Dueling Deep Q-Network και διάσπαση της συνάρτησης δράσης-αξίας σε συνιστώσες αξίας κατάστασης και πλεονεκτήματος ενεργειών

Συμπερασματικά, η αρχιτεκτονική Dueling DQN προσφέρει ένα ισχυρό, σταθερό και τεχνικά ώριμο πλαίσιο για την ανάπτυξη στρατηγικών αλγοριθμικής διαπραγμάτευσης.

4.3.3 Policy Network, Target Network και μηχανισμός βελτιστοποίησης

Η αποτελεσματικότητα της υλοποίησης του πράκτορα Dueling Deep Q-Network εξαρτάται καθοριστικά από τη συνεργασία του Policy Network, του Target Network, της Replay Memory και του μηχανισμού βελτιστοποίησης.

Το Policy Network αποτελεί το κύριο εκπαιδευόμενο νευρωνικό δίκτυο του συστήματος και είναι υπεύθυνο για την επεξεργασία των state vectors, τον υπολογισμό των Q-values και την επιλογή ενεργειών. Για κάθε κατάσταση s, το δίκτυο εκτιμά:

$$Q_{policy}(s, a; \theta) \qquad (4.11)$$

όπου:

Q_policy(s,a;θ): εκτιμώμενη τιμή της συνάρτησης δράσης-αξίας που υπολογίζεται από το Policy Network,

s: τρέχουσα κατάσταση (state) του περιβάλλοντος συναλλαγών,

a: εξεταζόμενη ενέργεια από τον διακριτό χώρο ενεργειών {Buy, Sell, Hold},

θ: σύνολο παραμέτρων (βάρη και μετατοπίσεις) του Policy Network.

Κατά τη διάρκεια της εκπαίδευσης, το policy network ενημερώνεται συνεχώς μέσω gradient descent, προσαρμόζοντας σταδιακά τα βάρη του.

Κατά τη διάρκεια της εκπαίδευσης, το policy network ενημερώνεται συνεχώς μέσω gradient descent, προσαρμόζοντας σταδιακά τα βάρη του.

Το Target Network αποτελεί αντίγραφο του policy network, το οποίο ενημερώνεται σε τακτά χρονικά διαστήματα και χρησιμοποιείται για τον υπολογισμό σταθερότερων τιμών-στόχων. Οι παράμετροί του συμβολίζονται ως θ^- και ενημερώνονται ως εξής:

$$\theta^- \leftarrow \theta \qquad (4.12)$$

όπου θ^- συμβολίζει τις παραμέτρους του target network.

Η ύπαρξη ξεχωριστού target network μειώνει σημαντικά την αστάθεια της εκπαίδευσης.

Κατά τη διαδικασία εκπαίδευσης, ο στόχος υπολογίζεται μέσω της εξίσωσης Bellman, ενώ σε περίπτωση τερματισμού επεισοδίου:

$$y = r \qquad (4.13)$$

Η απόκλιση μεταξύ πρόβλεψης και στόχου εκφράζεται μέσω της συνάρτησης απώλειας:

$$L(\theta) = (\text{Q\_policy}(\text{s}, \text{a}; \theta) - y)\text{^}2$$

Η ελαχιστοποίηση της συνάρτησης αυτής οδηγεί στη βελτιστοποίηση των παραμέτρων του policy network.

Η Replay Memory λειτουργεί ως μηχανισμός αποθήκευσης εμπειριών της μορφής (s, a, r, s′, done), επιτρέποντας τυχαία δειγματοληψία mini-batches και βελτιώνοντας τη σταθερότητα της εκπαίδευσης.

Η διαδικασία βελτιστοποίησης ακολουθεί επαναληπτικό κύκλο:

• δειγματοληψία mini-batch από replay memory

• υπολογισμός Bellman targets

• εκτίμηση Q_policy(s,a;θ)

• υπολογισμός loss

• εφαρμογή gradient descent

• ενημέρωση policy network

• περιοδική ενημέρωση target network

Η διαδικασία αυτή εφαρμόζεται επαναληπτικά κατά τη διάρκεια της εκπαίδευσης και ενσωματώνεται στη rolling-window ή walk-forward πειραματική δομή.

Σημαντικό είναι ότι η δομή policy/target/replay/optimization παραμένει κοινή τόσο στο BASE όσο και στο VWAP σενάριο.

Συμπερασματικά, η συνεργασία των παραπάνω μηχανισμών αποτελεί τον βασικό πυρήνα της μαθησιακής διαδικασίας του πράκτορα Dueling DQN.

4.3.4 Υλοποίηση του Trading Environment

Το Trading Environment αποτελεί τον βασικό μηχανισμό αλληλεπίδρασης μεταξύ του πράκτορα ενισχυτικής μάθησης και της χρηματοοικονομικής αγοράς, καθώς προσομοιώνει τη

δυναμική λειτουργία του περιβάλλοντος συναλλαγών και καθορίζει τον τρόπο με τον οποίο οι ενέργειες του πράκτορα μεταφράζονται σε πραγματικές μεταβολές του χαρτοφυλακίου και σε μαθησιακά σήματα ανταμοιβής. Η υλοποίησή του είναι καθοριστικής σημασίας, διότι συνδέει τη θεωρητική αρχιτεκτονική του Dueling Deep Q-Network με την πρακτική λειτουργία της αλγοριθμικής διαπραγμάτευσης.

Το περιβάλλον οργανώνεται σε διακριτά επεισόδια, όπου κάθε επεισόδιο αντιστοιχεί σε συγκεκριμένη χρονική περίοδο συναλλαγών, σύμφωνα με το εκάστοτε πειραματικό σενάριο. Κατά την έναρξη κάθε επεισοδίου πραγματοποιείται διαδικασία επαναφοράς (reset), κατά την οποία αρχικοποιούνται οι μεταβλητές του περιβάλλοντος, μηδενίζονται οι ανοικτές θέσεις, επαναφέρεται η αξία του χαρτοφυλακίου και δημιουργείται η αρχική κατάσταση του πράκτορα.

Κατά την έναρξη κάθε επεισοδίου, το trading environment αρχικοποιείται με διαθέσιμο κεφάλαιο 100.000 δολαρίων ΗΠΑ. Με την ολοκλήρωση κάθε επεισοδίου ή περιόδου αξιολόγησης, όλες οι ανοικτές θέσεις μηδενίζονται και το χαρτοφυλάκιο επανέρχεται στην αρχική του κατάσταση, ώστε να εξασφαλίζεται ανεξάρτητη αξιολόγηση μεταξύ διαφορετικών χρονικών περιόδων και πειραματικών φάσεων.

Η βασική λειτουργία του περιβάλλοντος εκτελείται μέσω της διαδικασίας step(action), η οποία εφαρμόζεται σε κάθε χρονικό βήμα. Σε κάθε βήμα:

• λαμβάνεται η επιλεγμένη ενέργεια του πράκτορα,

• ελέγχεται η εγκυρότητα της ενέργειας βάσει της τρέχουσας θέσης,

• εκτελείται αγορά, πώληση ή διακράτηση,

• ενημερώνεται η κατάσταση θέσης,

• υπολογίζεται η νέα αξία χαρτοφυλακίου,

• υπολογίζεται η ανταμοιβή, η οποία συνδυάζει στοιχεία απόδοσης και διαχείρισης κινδύνου, όπως περιγράφεται αναλυτικά στη συνάρτηση ανταμοιβής,

• δημιουργείται η νέα κατάσταση,

• ελέγχεται η ολοκλήρωση του επεισοδίου.

Η αναπαράσταση της κατάστασης βασίζεται στα προεπεξεργασμένα χαρακτηριστικά αγοράς, τεχνικούς δείκτες, μεταβλητές χαρτοφυλακίου και, στην εκδοχή VWAP, πρόσθετα

χαρακτηριστικά σχετιζόμενα με τον VWAP. Το περιβάλλον διασφαλίζει ότι κάθε νέα κατάσταση παρέχεται στον πράκτορα με χρονικά συνεπή τρόπο, αποφεύγοντας χρήση μελλοντικής πληροφορίας. Η χρονική αυτή συνέπεια είναι ιδιαίτερα σημαντική, καθώς αποτρέπει τη χρήση μελλοντικής πληροφορίας (look-ahead bias) και επιτρέπει πιο ρεαλιστική προσομοίωση πραγματικών συνθηκών διαπραγμάτευσης.

Ιδιαίτερη σημασία έχει η διαχείριση του χαρτοφυλακίου, καθώς το Trading Environment ενημερώνει διαρκώς την τρέχουσα θέση του πράκτορα, την τιμή εισόδου κάθε συναλλαγής, τα πραγματοποιημένα κέρδη ή ζημίες, τη συνολική αξία του χαρτοφυλακίου, καθώς και το πλήρες ιστορικό συναλλαγών. Η συνεχής αυτή ενημέρωση επιτρέπει τη ρεαλιστική αποτύπωση της επενδυτικής δραστηριότητας και διασφαλίζει ότι κάθε απόφαση του πράκτορα αξιολογείται με βάση τις πραγματικές οικονομικές της συνέπειες. Με τον τρόπο αυτό, η μαθησιακή διαδικασία δεν περιορίζεται σε θεωρητική πρόβλεψη τιμών, αλλά συνδέεται άμεσα με την πρακτική απόδοση μιας ολοκληρωμένης στρατηγικής διαπραγμάτευσης.

Το περιβάλλον υποστηρίζει επίσης τη λογική rolling-window / walk-forward, επιτρέποντας τη διαδοχική εκπαίδευση και αξιολόγηση του πράκτορα σε διαφορετικές χρονικές περιόδους. Στις φάσεις αξιολόγησης και τελικού ελέγχου (evaluation και test), η διαδικασία μάθησης απενεργοποιείται και ο πράκτορας εφαρμόζει αποκλειστικά την ήδη μαθημένη πολιτική, ώστε να αξιολογείται η δυνατότητα γενίκευσης του μοντέλου σε μη παρατηρημένα δεδομένα αγοράς.

Στο πλαίσιο της συγκριτικής αξιολόγησης BASE και VWAP, το Trading Environment παραμένει λειτουργικά αμετάβλητο. Η μοναδική διαφοροποίηση αφορά το σύνολο χαρακτηριστικών που συνθέτουν το state vector, επιτρέποντας μεθοδολογικά αυστηρότερη αποτίμηση της πραγματικής συμβολής του VWAP.

Η συνολική λειτουργία του περιβάλλοντος και η διασύνδεσή του με τη διαδικασία εκπαίδευσης μπορεί να συνοψιστεί ως εξής:

• αρχικοποίηση επεισοδίου,

• παραγωγή αρχικής κατάστασης,

• λήψη ενέργειας,

• εφαρμογή συναλλαγής,

• ενημέρωση χαρτοφυλακίου,

• υπολογισμός ανταμοιβής,

• μετάβαση σε νέα κατάσταση,

• αποθήκευση εμπειρίας,

• επανάληψη έως ολοκλήρωση επεισοδίου.

```
1. Λήψη τρέχουσας κατάστασης s
2. Λήψη επιλεγμένης ενέργειας a
3. Αν a = Buy:
   3.1 Άνοιγμα ή ενίσχυση long θέσης
   3.2 Υπολογισμός κόστους συναλλαγής TC_t
4. Αν a = Sell:
   4.1 Άνοιγμα ή ενίσχυση short θέσης
   4.2 Υπολογισμός κόστους συναλλαγής TC_t
5. Αν a = Hold:
   5.1 Διατήρηση υφιστάμενης θέσης
6. Υπολογισμός νέας αξίας χαρτοφυλακίου V_t
7. Υπολογισμός βασικής ανταμοιβής:
   r_base(t) = r_pnl(t) - TC_t
8. Υπολογισμός δείκτη Sortino:
   SR_t = calculate_sortino_ratio()
9. Υπολογισμός drawdown:
   DD_t = calculate_drawdown()
10.   Υπολογισμός συνολικής ανταμοιβής:
   r_t = α · r_base(t) + β · SR_t − γ · DD_t
11.   Περιορισμός ανταμοιβής στο επιτρεπτό εύρος:
   r_t = clip(r_t, −10, 10)
12.   Μετάβαση στο επόμενο χρονικό βήμα
13.   Δημιουργία νέας κατάστασης s'
14.   Έλεγχος ολοκλήρωσης επεισοδίου:
   Αν τέλος δεδομένων:
   done = True
   Αλλιώς:
   done = False
15.   Επιστροφή:
   (s', r_t, done)
```

Ψευδοκώδικας 3 Λειτουργία Trading Environment (step function)

4.3.5 Υλοποίηση της Reward Function

Η συνάρτηση ανταμοιβής (Reward Function) αποτελεί έναν από τους σημαντικότερους μηχανισμούς του συστήματος ενισχυτικής μάθησης, καθώς μετατρέπει τις συνέπειες των ενεργειών του πράκτορα σε μαθησιακό σήμα, το οποίο χρησιμοποιείται για την προσαρμογή της πολιτικής λήψης αποφάσεων. Μέσω της reward function, ο πράκτορας αξιολογεί τη χρηματοοικονομική αποτελεσματικότητα των ενεργειών του και μαθαίνει σταδιακά να επιλέγει στρατηγικές που οδηγούν σε υψηλότερη μακροπρόθεσμη απόδοση.

Στην παρούσα εργασία, η συνάρτηση ανταμοιβής δεν βασίζεται αποκλειστικά στη μεταβολή της αξίας του χαρτοφυλακίου, αλλά συνδυάζει στοιχεία απόδοσης και διαχείρισης κινδύνου. Η μαθηματική διατύπωση της βασικής ανταμοιβής και της τελικής συνάρτησης ανταμοιβής παρουσιάστηκε αναλυτικά στην Ενότητα 3.3 και αποτυπώνεται στις εξισώσεις (4.6) και (4.9).

Η υλοποίηση της reward function πραγματοποιείται στο εσωτερικό της διαδικασίας step() του Trading Environment και εκτελείται μετά από κάθε εφαρμογή ενέργειας αγοράς, πώλησης ή διακράτησης. Αρχικά υπολογίζεται η συνιστώσα απόδοσης του χαρτοφυλακίου, λαμβάνοντας υπόψη το οικονομικό αποτέλεσμα της θέσης και το αντίστοιχο κόστος συναλλαγής. Στη συνέχεια υπολογίζονται οι συνιστώσες που σχετίζονται με την προσαρμογή της απόδοσης ως προς τον κίνδυνο και με την επιβολή ποινής σε περιόδους αυξημένου drawdown.

Ο δείκτης Sortino χρησιμοποιείται ως μέτρο προσαρμοσμένης ως προς τον κίνδυνο απόδοσης, καθώς εστιάζει αποκλειστικά στις αρνητικές αποκλίσεις των αποδόσεων. Με τον τρόπο αυτό, ο πράκτορας επιβραβεύεται όχι μόνο για την επίτευξη κερδών αλλά και για τη διατήρηση ευνοϊκού λόγου απόδοσης προς καθοδικό κίνδυνο.

Παράλληλα, η συνιστώσα drawdown λειτουργεί ως μηχανισμός ποινής για σημαντικές μειώσεις της αξίας του χαρτοφυλακίου σε σχέση με προηγούμενα υψηλά επίπεδα. Η ενσωμάτωση του συγκεκριμένου όρου συμβάλλει στη διαμόρφωση πιο σταθερών στρατηγικών και αποθαρρύνει συμπεριφορές που οδηγούν σε υπερβολική ανάληψη κινδύνου.

Για λόγους αριθμητικής σταθερότητας κατά την εκπαίδευση του νευρωνικού δικτύου, η τελική ανταμοιβή περιορίζεται σε προκαθορισμένο εύρος τιμών μέσω διαδικασίας clipping. Η πρακτική αυτή αποτρέπει την εμφάνιση ακραίων τιμών ανταμοιβής, οι οποίες θα μπορούσαν να επηρεάσουν αρνητικά τη σταθερότητα της μαθησιακής διαδικασίας και τη σύγκλιση του μοντέλου.

Η υπολογισμένη ανταμοιβή επιστρέφεται στον πράκτορα ως μέρος της εμπειρίας αλληλεπίδρασης με το περιβάλλον και αποθηκεύεται στη replay memory μαζί με την τρέχουσα κατάσταση, την επιλεγμένη ενέργεια, την επόμενη κατάσταση και την ένδειξη ολοκλήρωσης επεισοδίου. Με τον τρόπο αυτό, η reward function συνδέεται άμεσα με τη διαδικασία εκμάθησης μέσω της εξίσωσης Bellman και επηρεάζει τη σταδιακή προσαρμογή της συνάρτησης αξίας δράσης του πράκτορα.

Ιδιαίτερα σημαντικό είναι ότι η ενσωμάτωση του δείκτη VWAP δεν μεταβάλλει τη reward function. Ο VWAP χρησιμοποιείται αποκλειστικά ως πρόσθετο χαρακτηριστικό εισόδου στο διάνυσμα κατάστασης (state vector), ενώ ο μηχανισμός υπολογισμού της ανταμοιβής παραμένει αμετάβλητος τόσο στην εκδοχή BASE όσο και στην εκδοχή VWAP. Η επιλογή αυτή διασφαλίζει ότι τυχόν διαφοροποιήσεις στην απόδοση μπορούν να αποδοθούν στην πληροφοριακή συνεισφορά του VWAP και όχι σε αλλαγές του μαθησιακού σήματος.

Η υλοποίηση αυτή επιτρέπει την ταυτόχρονη αξιολόγηση απόδοσης και κινδύνου, συνδέοντας τη μαθησιακή διαδικασία με περισσότερο ρεαλιστικά κριτήρια χρηματοοικονομικής διαχείρισης σε σχέση με μια reward function που βασίζεται αποκλειστικά στη μεταβολή της αξίας του χαρτοφυλακίου.

```
1. Λήψη προηγούμενης αξίας χαρτοφυλακίου V_(t-1)
2. Λήψη τρέχουσας αξίας χαρτοφυλακίου V_t
3. Υπολογισμός συνιστώσας PnL:
   r_pnl(t)
4. Υπολογισμός κόστους συναλλαγών:
   TC_t
5. Υπολογισμός βασικής ανταμοιβής:
   r_base(t) = r_pnl(t) - TC_t
6. Υπολογισμός Sortino Ratio:
   SR_t
7. Υπολογισμός drawdown:
   DD_t
8. Υπολογισμός συνολικής ανταμοιβής:
   r_t = α·r_base(t) + β·SR_t - γ·DD_t
9. Περιορισμός ανταμοιβής:
   r_t = clip(r_t, -10, 10)
10.   Επιστροφή ανταμοιβής:
   r_t
```

Ψευδοκώδικας 4 Υπολογισμός ανταμοιβής (Reward Function)

4.4 Περιγραφή των υλοποιήσεων των πειραματικών σεναρίων

4.4.1 Υλοποίηση του Pilot Walk-Forward Scenario (6-Day Dataset)

Το πρώτο πειραματικό σενάριο της παρούσας εργασίας βασίζεται στο Pilot Walk-Forward Scenario (6-Day Dataset) και χρησιμοποιείται ως αρχικό πειραματικό πλαίσιο περιορισμένης χρονικής κλίμακας για την αξιολόγηση της λειτουργικότητας, της τεχνικής ορθότητας και της βασικής αποδοτικότητας του πράκτορα Dueling Deep Q-Network. Ο κύριος στόχος του συγκεκριμένου σεναρίου δεν είναι η διερεύνηση της μακροχρόνιας ικανότητας γενίκευσης του πράκτορα, αλλά η πιλοτική επιβεβαίωση της ορθής λειτουργίας του συνολικού συστήματος και η αρχική αποτίμηση της ικανότητάς του να μαθαίνει αποδοτικές στρατηγικές συναλλαγών σε ενδοημερήσιο περιβάλλον.

Το dataset περιλαμβάνει ιστορικά χρηματοοικονομικά δεδομένα υψηλής συχνότητας (minute-level) για έξι διαδοχικές χρηματιστηριακές συνεδριάσεις. Κάθε χρονικό βήμα αντιστοιχεί σε ένα λεπτό διαπραγμάτευσης, επιτρέποντας στο περιβάλλον συναλλαγών να προσομοιώνει ρεαλιστικές συνθήκες ενδοημερήσιας αγοράς. Για κάθε εξεταζόμενη μετοχή χρησιμοποιούνται ως βασικά χαρακτηριστικά οι τιμές Open, High, Low, Close και Volume, καθώς και οι τεχνικοί δείκτες που προκύπτουν από τη διαδικασία προεπεξεργασίας (Ενότητα 3.5). Στην εκδοχή VWAP προστίθενται επιπλέον χαρακτηριστικά που σχετίζονται με τον δείκτη VWAP, χωρίς να μεταβάλλεται οποιοδήποτε άλλο στοιχείο της αρχιτεκτονικής του συστήματος.

Η διαδικασία εκπαίδευσης οργανώνεται σε πέντε διαδοχικά επεισόδια. Κάθε επεισόδιο αξιοποιεί τις ίδιες έξι διαδοχικές συνεδριάσεις του συνόλου δεδομένων και περιλαμβάνει πέντε διαδοχικούς κύκλους εκπαίδευσης και συναλλαγών (Train → Trade). Σε κάθε κύκλο ο πράκτορας εκπαιδεύεται στα δεδομένα μίας χρηματιστηριακής συνεδρίασης (Day D) και στη συνέχεια εφαρμόζει την πολιτική που έχει μάθει στην αμέσως επόμενη συνεδρίαση (Day D+1), χωρίς να πραγματοποιείται περαιτέρω ενημέρωση των παραμέτρων του μοντέλου κατά τη διάρκεια της φάσης συναλλαγών. Με την ολοκλήρωση κάθε επεισοδίου το περιβάλλον επανεκκινείται και η διαδικασία επαναλαμβάνεται από την αρχή, επιτρέποντας την ανεξάρτητη αξιολόγηση πολλαπλών εκτελέσεων του ίδιου πειραματικού σεναρίου.

Η διαδικασία αυτή αποτελεί μία απλή μορφή walk-forward εκτέλεσης, κατά την οποία η εκπαίδευση προηγείται πάντοτε χρονικά της εφαρμογής της στρατηγικής, διατηρώντας τη φυσική χρονική ακολουθία των δεδομένων και αποφεύγοντας τη χρήση μελλοντικής πληροφορίας (look-ahead bias). Το χαρτοφυλάκιο αρχικοποιείται μία μόνο φορά με αρχικό

κεφάλαιο 100.000 USD και εξελίσσεται διαδοχικά κατά τη διάρκεια και των πέντε κύκλων Train → Trade, χωρίς επανεκκίνηση μεταξύ τους. Ως τελική αξία του χαρτοφυλακίου καταγράφεται η αξία που προκύπτει μετά την ολοκλήρωση της πέμπτης και τελευταίας trading day του επεισοδίου.

Στο πλαίσιο αυτό υλοποιούνται δύο διακριτές πειραματικές εκδοχές:

- η βασική εκδοχή (BASE), όπου το state vector περιλαμβάνει χαρακτηριστικά αγοράς, τεχνικούς δείκτες και μεταβλητές κατάστασης του χαρτοφυλακίου,
- η εκδοχή VWAP, όπου στο παραπάνω σύνολο χαρακτηριστικών προστίθενται και χαρακτηριστικά που σχετίζονται με τον δείκτη VWAP.

Η διαφοροποίηση αυτή αφορά αποκλειστικά την αναπαράσταση της κατάστασης (state representation) και όχι τον μηχανισμό εκτέλεσης συναλλαγών, τη συνάρτηση ανταμοιβής (Reward Function), τη δομή του Trading Environment, την αρχιτεκτονική του Dueling Deep Q-Network ή τη διαδικασία εκπαίδευσης του πράκτορα. Συνεπώς, και στις δύο εκδοχές εφαρμόζονται οι ίδιες ακριβώς συνθήκες εκπαίδευσης και αξιολόγησης, με μοναδική μεταβλητή τη σύνθεση του state vector. Με τον τρόπο αυτό διασφαλίζεται ότι οποιαδήποτε διαφοροποίηση στη συμπεριφορά ή στην απόδοση του πράκτορα μπορεί να αποδοθεί αποκλειστικά στην πληροφοριακή συνεισφορά των χαρακτηριστικών που προέρχονται από τον δείκτη VWAP και όχι σε μεταβολές της αρχιτεκτονικής ή της μαθησιακής διαδικασίας.

Η διαδικασία εκπαίδευσης ακολουθεί τη μεθοδολογία που παρουσιάστηκε στις Ενότητες 3.6 και 3.7 και περιλαμβάνει τη φόρτωση των δεδομένων, την προεπεξεργασία και κανονικοποίησή τους, τη δημιουργία των state vectors, τη διαδοχική αλληλεπίδραση πράκτορα–περιβάλλοντος κατά τη διάρκεια του επεισοδίου, την αποθήκευση εμπειριών στη replay memory και την ενημέρωση του policy network μέσω της διαδικασίας βελτιστοποίησης. Το περιορισμένο χρονικό εύρος του σεναρίου επιτρέπει ταχεία εκπαίδευση και διευκολύνει τον εντοπισμό τεχνικών προβλημάτων ή πιθανών ασταθειών της υλοποίησης.

Η αξιολόγηση πραγματοποιείται διαδοχικά μετά από κάθε φάση εκπαίδευσης, εφαρμόζοντας τη μαθημένη πολιτική στην αμέσως επόμενη χρηματιστηριακή συνεδρίαση χωρίς περαιτέρω ενημέρωση των παραμέτρων του μοντέλου. Κατά τη διάρκεια των πέντε κύκλων Train → Trade καταγράφονται οι συναλλαγές, η εξέλιξη της αξίας του χαρτοφυλακίου, οι σωρευτικές

ανταμοιβές και οι βασικές μετρικές απόδοσης, οι οποίες χρησιμοποιούνται ως αρχικό σημείο αναφοράς για τη σύγκριση μεταξύ των εκδοχών BASE και VWAP.

Ο ερευνητικός ρόλος του συγκεκριμένου σεναρίου είναι ιδιαίτερα σημαντικός, καθώς λειτουργεί ως προκαταρκτικό στάδιο επαλήθευσης της συνολικής μεθοδολογίας. Παρέχει ένα ελεγχόμενο περιβάλλον χαμηλού υπολογιστικού κόστους, διευκολύνει την κατανόηση της συμπεριφοράς του πράκτορα και επιτρέπει την αρχική εμπειρική διερεύνηση της επίδρασης της ενσωμάτωσης των χαρακτηριστικών VWAP.

Παρά τα πλεονεκτήματα αυτά, το περιορισμένο χρονικό εύρος του σεναρίου εισάγει και περιορισμούς, όπως η μειωμένη δυνατότητα γενίκευσης των αποτελεσμάτων και η αυξημένη επίδραση της βραχυπρόθεσμης μεταβλητότητας της αγοράς.

Συνολικά, το Pilot Walk-Forward Scenario (6-Day Dataset) αποτελεί το πρώτο στάδιο της πειραματικής διαδικασίας, προσφέροντας ένα ελεγχόμενο πλαίσιο επαλήθευσης της υλοποίησης και της αρχιτεκτονικής Dueling DQN, καθώς και μία πρώτη εμπειρική αξιολόγηση της συμβολής της ενσωμάτωσης των χαρακτηριστικών VWAP στη διαδικασία λήψης αποφάσεων του πράκτορα.

4.4.2 Υλοποίηση του εκτεταμένου walk-forward πειραματικού σεναρίου (111 συνεδριάσεις)

Το δεύτερο πειραματικό σενάριο της παρούσας εργασίας αποτελεί το κύριο πλαίσιο αξιολόγησης του προτεινόμενου συστήματος ενισχυτικής μάθησης και βασίζεται σε σύνολο 111 διαδοχικών χρηματιστηριακών συνεδριάσεων. Σε αντίθεση με το πιλοτικό σενάριο περιορισμένης κλίμακας (6-Day dataset), το συγκεκριμένο σενάριο χρησιμοποιεί σημαντικά μεγαλύτερο χρονικό εύρος δεδομένων και αποσκοπεί στην αξιολόγηση της συμπεριφοράς του πράκτορα σε εκτεταμένο χρονικό ορίζοντα, επιτρέποντας την εξαγωγή περισσότερο αξιόπιστων και στατιστικά ισχυρών συμπερασμάτων.

Το dataset περιλαμβάνει ιστορικά ενδοημερήσια δεδομένα υψηλής συχνότητας (minute-level), όπου κάθε χρονικό βήμα αντιστοιχεί σε ένα λεπτό διαπραγμάτευσης. Για κάθε εξεταζόμενη μετοχή χρησιμοποιούνται οι τιμές Open, High, Low, Close και Volume, καθώς και οι τεχνικοί δείκτες που παράγονται κατά τη διαδικασία προεπεξεργασίας (Ενότητα 3.5). Στην εκδοχή VWAP προστίθενται δύο επιπλέον χαρακτηριστικά που προέρχονται από τον δείκτη VWAP, χωρίς να μεταβάλλεται οποιοδήποτε άλλο στοιχείο της αρχιτεκτονικής.

Η διαδικασία εκτέλεσης οργανώνεται σε διαδοχικά ανεξάρτητα πενθήμερα επεισόδια (episodes), τα οποία ακολουθούν λογική walk-forward. Κάθε επεισόδιο αποτελείται από πέντε διαδοχικούς κύκλους εκπαίδευσης και εφαρμογής της στρατηγικής (Train → Trade). Σε κάθε κύκλο ο πράκτορας εκπαιδεύεται στα δεδομένα μίας χρηματιστηριακής συνεδρίασης και στη συνέχεια εφαρμόζει την εκμαθημένη πολιτική στην αμέσως επόμενη διαθέσιμη συνεδρίαση, χωρίς να πραγματοποιείται περαιτέρω ενημέρωση των παραμέτρων του μοντέλου κατά τη διάρκεια της φάσης συναλλαγών. Με την έναρξη κάθε νέου επεισοδίου το χαρτοφυλάκιο επαναφέρεται στην αρχική αξία των 100.000 δολαρίων και όλες οι ανοικτές θέσεις μηδενίζονται, ώστε κάθε επεισόδιο να αξιολογείται ανεξάρτητα από τα προηγούμενα. Το σημείο εκκίνησης μετατοπίζεται στη συνέχεια κατά πέντε χρηματιστηριακές συνεδριάσεις και η ίδια διαδικασία επαναλαμβάνεται μέχρι την εξάντληση των διαθέσιμων δεδομένων

Η προσέγγιση αυτή διατηρεί τη χρονική ακολουθία των δεδομένων και αποφεύγει τη χρήση μελλοντικής πληροφορίας (look-ahead bias), προσεγγίζοντας με περισσότερο ρεαλιστικό τρόπο τη λειτουργία ενός συστήματος αλγοριθμικής διαπραγμάτευσης σε πραγματικές συνθήκες αγοράς.

Όπως και στο προηγούμενο πειραματικό σενάριο, υλοποιούνται δύο διακριτές εκδοχές του συστήματος:

• η βασική εκδοχή (BASE), όπου το state vector περιλαμβάνει χαρακτηριστικά αγοράς, τεχνικούς δείκτες και μεταβλητές κατάστασης χαρτοφυλακίου,

• η εκδοχή VWAP, όπου στο παραπάνω σύνολο χαρακτηριστικών προστίθενται δύο επιπλέον χαρακτηριστικά που σχετίζονται με τον δείκτη VWAP.

Η διαφοροποίηση αυτή αφορά αποκλειστικά την αναπαράσταση της κατάστασης (state representation) και όχι τον μηχανισμό εκτέλεσης συναλλαγών, τη συνάρτηση ανταμοιβής (Reward Function), τη δομή του Trading Environment, την αρχιτεκτονική του Dueling Deep Q-Network ή τη διαδικασία εκπαίδευσης του πράκτορα. Συνεπώς, και στις δύο εκδοχές εφαρμόζονται οι ίδιες ακριβώς συνθήκες εκπαίδευσης και αξιολόγησης, με μοναδική μεταβλητή τη σύνθεση του state vector. Με τον τρόπο αυτό διασφαλίζεται ότι οποιαδήποτε διαφοροποίηση στη συμπεριφορά ή στην απόδοση του πράκτορα μπορεί να αποδοθεί αποκλειστικά στην πληροφοριακή συνεισφορά των χαρακτηριστικών που προέρχονται από τον δείκτη VWAP.

Η διαδικασία εκπαίδευσης είναι σημαντικά εκτενέστερη σε σχέση με το Pilot Walk-Forward Scenario, καθώς αξιοποιείται πολύ μεγαλύτερο πλήθος χρηματιστηριακών συνεδριάσεων, εκτελούνται περισσότερα ανεξάρτητα επεισόδια και, συνολικά, πραγματοποιούνται περισσότεροι κύκλοι εκπαίδευσης και συναλλαγών. Ως αποτέλεσμα, παράγεται σημαντικά μεγαλύτερος αριθμός εμπειριών (experiences), πραγματοποιούνται περισσότερες ενημερώσεις του policy network μέσω της replay memory και ο πράκτορας εκπαιδεύεται σε ευρύτερο φάσμα συνθηκών αγοράς, γεγονός που ευνοεί τη σταθερότητα της μαθημένης πολιτικής.

Καθ’ όλη τη διάρκεια της εκτέλεσης καταγράφονται αναλυτικά οι συναλλαγές, η εξέλιξη της αξίας του χαρτοφυλακίου, οι ημερήσιες και σωρευτικές ανταμοιβές, ο αριθμός των long και short συναλλαγών, το κόστος συναλλαγών, καθώς και πρόσθετες λειτουργικές μετρικές, όπως ο μέγιστος αριθμός ανοικτών θέσεων, η μέγιστη έκθεση του χαρτοφυλακίου και ο συνολικός όγκος συναλλαγών. Με την ολοκλήρωση της εκτέλεσης, τα αποτελέσματα αποθηκεύονται αυτόματα σε αρχεία CSV, τα οποία χρησιμοποιούνται για τη συγκριτική ανάλυση μεταξύ των εκδοχών BASE και VWAP και για την παραγωγή των πινάκων και των γραφημάτων που παρουσιάζονται στο επόμενο κεφάλαιο.

Το συγκεκριμένο σενάριο αποτελεί τον κύριο ερευνητικό άξονα της εργασίας και το βασικό πειραματικό πλαίσιο εξαγωγής των τελικών αποτελεσμάτων. Το μεγαλύτερο χρονικό εύρος των δεδομένων μειώνει την επίδραση βραχυπρόθεσμων διακυμάνσεων της αγοράς, ενισχύει τη στατιστική αξιοπιστία των αποτελεσμάτων και επιτρέπει ουσιαστικότερη αξιολόγηση της συμπεριφοράς του πράκτορα σε διαφορετικές συνθήκες αγοράς.

Τα βασικά πλεονεκτήματα του σεναρίου περιλαμβάνουν:

• μεγαλύτερο χρονικό εύρος δεδομένων,

• αυξημένη στατιστική αξιοπιστία,

• πληρέστερη αξιολόγηση της συμπεριφοράς του πράκτορα,

• ρεαλιστικότερη προσομοίωση συνθηκών αλγοριθμικής διαπραγμάτευσης.

Ωστόσο, το εκτεταμένο αυτό πειραματικό σενάριο συνοδεύεται από αυξημένο υπολογιστικό κόστος, μεγαλύτερη διάρκεια εκτέλεσης και σημαντικά αυξημένο όγκο παραγόμενων δεδομένων προς ανάλυση.

Συνολικά, το εκτεταμένο walk-forward πειραματικό σενάριο αποτελεί το κύριο περιβάλλον αξιολόγησης της παρούσας εργασίας. Η εφαρμογή της ίδιας μεθοδολογίας στις δύο εκδοχές του συστήματος (BASE και VWAP), σε συνδυασμό με το μεγάλο χρονικό εύρος του dataset, επιτρέπει τη δίκαιη και αξιόπιστη διερεύνηση της συνεισφοράς του δείκτη VWAP στην απόδοση του πράκτορα Dueling Deep Q-Network.

4.4.3 Διαφορές στη διαδικασία εκπαίδευσης και αξιολόγησης μεταξύ των πειραματικών σεναρίων

Τα δύο βασικά πειραματικά σενάρια της παρούσας εργασίας βασίζονται στην ίδια θεμελιώδη αρχιτεκτονική Dueling Deep Q-Network και εφαρμόζουν τον ίδιο μηχανισμό ενισχυτικής μάθησης. Επιπλέον, ακολουθούν την ίδια λογική walk-forward εκτέλεσης, κατά την οποία η εκπαίδευση προηγείται χρονικά της εφαρμογής της στρατηγικής (Train → Trade), διατηρώντας τη φυσική χρονική ακολουθία των δεδομένων και αποφεύγοντας τη χρήση μελλοντικής πληροφορίας (look-ahead bias). Ωστόσο, διαφοροποιούνται ως προς το μέγεθος του συνόλου δεδομένων, το πλήθος των κύκλων εκπαίδευσης και συναλλαγών που εκτελούνται, τον όγκο

των εμπειριών που παράγονται κατά την εκπαίδευση και τον λειτουργικό τους ρόλο στο πλαίσιο της πειραματικής διαδικασίας.

Το πρώτο σενάριο, Pilot Walk-Forward Scenario (6-Day Dataset), λειτουργεί ως πιλοτικό πειραματικό πλαίσιο περιορισμένης κλίμακας. Η μικρή χρονική του έκταση επιτρέπει ταχεία εκτέλεση, χαμηλό υπολογιστικό κόστος και αποτελεσματική τεχνική επαλήθευση της υλοποίησης. Το σενάριο αυτό χρησιμοποιείται κυρίως για:

• έλεγχο της ορθής λειτουργίας της αρχιτεκτονικής,

• επιβεβαίωση της λειτουργικότητας των επιμέρους υποσυστημάτων,

• προκαταρκτική αξιολόγηση της συμπεριφοράς του πράκτορα,

• αρχική σύγκριση μεταξύ των εκδοχών BASE και VWAP.

Αντίθετα, το Extended Walk-Forward Scenario (111-Day Dataset) αποτελεί το κύριο πειραματικό πλαίσιο της εργασίας και αξιοποιεί το πλήρες σύνολο των 111 χρηματιστηριακών συνεδριάσεων. Η ίδια λογική walk-forward εφαρμόζεται σε σημαντικά μεγαλύτερη χρονική κλίμακα, μέσω διαδοχικών ανεξάρτητων πενθήμερων επεισοδίων. Κάθε επεισόδιο ξεκινά εκ νέου με αρχική αξία χαρτοφυλακίου 100.000 δολαρίων και μηδενικές ανοικτές θέσεις, ενώ το σημείο εκκίνησης μετατοπίζεται κατά πέντε χρηματιστηριακές συνεδριάσεις. Με τον τρόπο αυτό αξιοποιείται το σύνολο των 111 συνεδριάσεων χωρίς να δημιουργείται εξάρτηση μεταξύ διαδοχικών επεισοδίων.

Η κύρια διαφορά μεταξύ των δύο σεναρίων έγκειται στην κλίμακα της εκτέλεσης. Στο σενάριο 6-Day ο πράκτορας εκπαιδεύεται και αξιολογείται σε περιορισμένο αριθμό συνεδριάσεων, γεγονός που επιτρέπει γρήγορη επαλήθευση της λειτουργικότητας του συστήματος. Αντίθετα, στο εκτεταμένο σενάριο χρησιμοποιείται σημαντικά μεγαλύτερος όγκος δεδομένων, εκτελούνται περισσότεροι κύκλοι εκπαίδευσης και συναλλαγών, παράγεται μεγαλύτερος αριθμός εμπειριών που αποθηκεύονται στη replay memory και πραγματοποιούνται περισσότερες ενημερώσεις του policy network. Η εκτενέστερη αυτή διαδικασία επιτρέπει την αξιολόγηση της συμπεριφοράς του πράκτορα σε μεγαλύτερο εύρος συνθηκών της αγοράς και οδηγεί στα αποτελέσματα που παρουσιάζονται στο Κεφάλαιο 5.

Αντίστοιχη διαφοροποίηση παρατηρείται και στη διαδικασία αξιολόγησης. Στο σενάριο 6-Day η αξιολόγηση χρησιμοποιείται κυρίως ως λειτουργική επιβεβαίωση της ορθής υλοποίησης και ως πρώτη ένδειξη της επίδρασης του VWAP. Αντίθετα, στο Extended Walk-Forward Scenario

η αξιολόγηση προκύπτει από το σύνολο των διαδοχικών κύκλων Train → Trade που εκτελούνται σε όλα τα επεισόδια. Οι μετρικές που συλλέγονται από κάθε συνεδρία συγκεντρώνονται και χρησιμοποιούνται για τη συγκριτική αποτίμηση των εκδοχών BASE και VWAP, καθώς και για την εξαγωγή των τελικών συμπερασμάτων της εργασίας.

Κοινό χαρακτηριστικό και των δύο σεναρίων αποτελεί η διατήρηση σταθερής της βασικής αρχιτεκτονικής και των μηχανισμών μάθησης. Η διαφοροποίηση μεταξύ των εκδοχών BASE και VWAP περιορίζεται αποκλειστικά στο σύνολο των χαρακτηριστικών που συνθέτουν το state vector και δεν επηρεάζει τη Reward Function, το Trading Environment, την αρχιτεκτονική του Dueling Deep Q-Network ή τη διαδικασία εκπαίδευσης. Η σχεδιαστική αυτή επιλογή διασφαλίζει μεθοδολογική συνέπεια και επιτρέπει η οποιαδήποτε διαφοροποίηση στα αποτελέσματα να αποδοθεί αποκλειστικά στην πληροφοριακή συνεισφορά του δείκτη VWAP.

Συνολικά, τα δύο πειραματικά σενάρια επιτελούν συμπληρωματικούς αλλά διακριτούς ρόλους. Το πρώτο λειτουργεί ως ελεγχόμενο περιβάλλον αρχικής επαλήθευσης της υλοποίησης, ενώ το δεύτερο αποτελεί το κύριο πειραματικό πλαίσιο από το οποίο προκύπτουν τα αποτελέσματα και τα τελικά επιστημονικά συμπεράσματα της εργασίας. Η εφαρμογή της ίδιας μεθοδολογίας walk-forward τόσο στο Pilot Walk-Forward Scenario όσο και στο Extended Walk-Forward Scenario εξασφαλίζει συνέπεια στον πειραματικό σχεδιασμό και επιτρέπει η σύγκριση μεταξύ των εκδοχών BASE και VWAP να πραγματοποιείται υπό κοινές συνθήκες εκπαίδευσης και αξιολόγησης, ενισχύοντας τη συνολική επιστημονική εγκυρότητα της εργασίας.

4.5 Διαθεσιμότητα Κώδικα

Ο πλήρης κώδικας της παρούσας εργασίας είναι διαθέσιμος σε δημόσιο αποθετήριο GitHub στον ακόλουθο σύνδεσμο:

https://github.com/athasch-glitch/duelling-dqn-intraday-trading

Το αποθετήριο περιλαμβάνει όλες τις υλοποιήσεις των αλγορίθμων, τα πειραματικά σενάρια και τα scripts επεξεργασίας δεδομένων, επιτρέποντας την αναπαραγωγή των αποτελεσμάτων.

5. Αποτελέσματα και Συγκριτική Αξιολόγηση

5.1 Αποτελέσματα του Pilot Walk-Forward Scenario (6-Day Dataset)

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα του πρώτου πειραματικού σεναρίου (Pilot Walk-Forward Scenario), το οποίο βασίζεται σε σύνολο δεδομένων έξι διαδοχικών χρηματιστηριακών συνεδριάσεων. Στόχος της ανάλυσης είναι η συγκριτική αξιολόγηση της πιθανής επίδρασης της ενσωμάτωσης του δείκτη VWAP στη μαθησιακή διαδικασία και στη στρατηγική συμπεριφορά του πράκτορα. Το συγκεκριμένο σενάριο λειτουργεί ως πιλοτικό πειραματικό πλαίσιο, επιτρέποντας την αρχική επαλήθευση της λειτουργίας του συστήματος και την πρώτη συγκριτική αξιολόγηση των εκδοχών BASE και VWAP.

Τα πειράματα πραγματοποιήθηκαν σε επιλεγμένο σύνολο μετοχών, εφαρμόζοντας κοινή αρχιτεκτονική Dueling Deep Q-Network, κοινό Trading Environment, ίδια Reward Function και ίδιους μηχανισμούς εκπαίδευσης και αξιολόγησης, ώστε η μοναδική ουσιαστική διαφοροποίηση να αφορά την παρουσία ή μη των χαρακτηριστικών VWAP στην αναπαράσταση της κατάστασης (state vector).

Η απόδοση κάθε εκδοχής αξιολογείται πρωτίστως μέσω της τελικής αξίας του χαρτοφυλακίου στο τέλος του πενθήμερου επεισοδίου, η οποία χρησιμοποιείται ως κύρια μετρική συγκριτικής αποτίμησης μεταξύ των εκδοχών BASE και VWAP.

Μετοχή	BASE	VWAP
CHK	107.808,12	110.342,70
AMD	152.667,46	134.445,43
AES	137.907,99	132.105,18
F	144.347,42	127.764,93

Πίνακας 1 Τελική αξία χαρτοφυλακίου στο Pilot Walk-Forward Scenario (6-Day Dataset)

Από τα αποτελέσματα προκύπτει ότι η ενσωμάτωση των VWAP χαρακτηριστικών δεν επιδρά με ενιαίο τρόπο σε όλες τις εξεταζόμενες μετοχές. Στην περίπτωση της CHK παρατηρείται βελτιωμένη τελική απόδοση υπέρ της VWAP εκδοχής, ενώ στις μετοχές AMD, AES και F η βασική εκδοχή εμφανίζει υψηλότερη τελική αξία χαρτοφυλακίου.

Η εικόνα αυτή υποδηλώνει ότι η προσθήκη VWAP χαρακτηριστικών δεν οδηγεί αυτομάτως σε γενικευμένη βελτίωση της στρατηγικής, αλλά η επίδρασή της εξαρτάται από τη δυναμική του εκάστοτε χρηματοοικονομικού περιβάλλοντος.

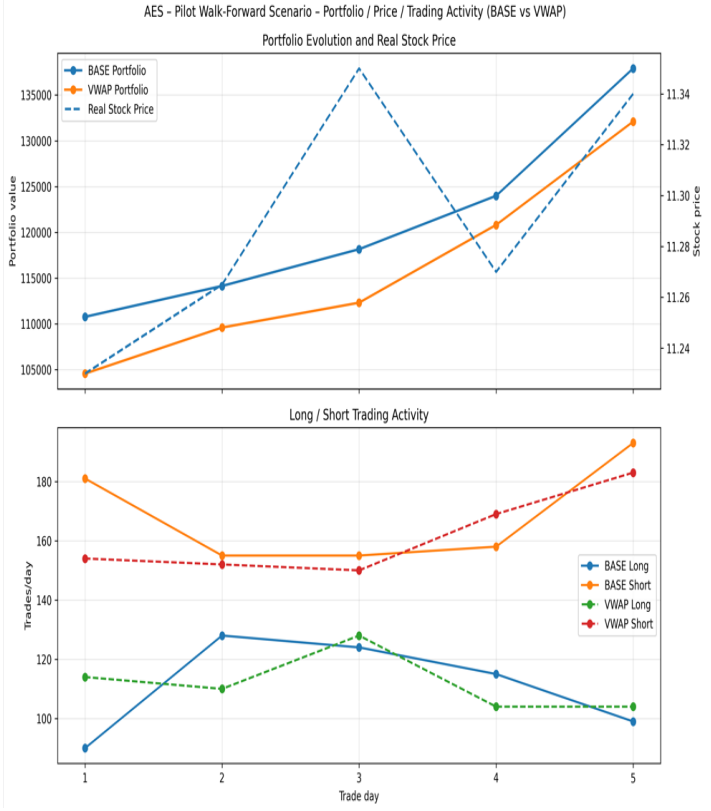
Για πληρέστερη αξιολόγηση, η ανάλυση δεν περιορίζεται μόνο στην τελική τιμή χαρτοφυλακίου, αλλά επεκτείνεται επίσης:

• στην εξέλιξη της αξίας του χαρτοφυλακίου,

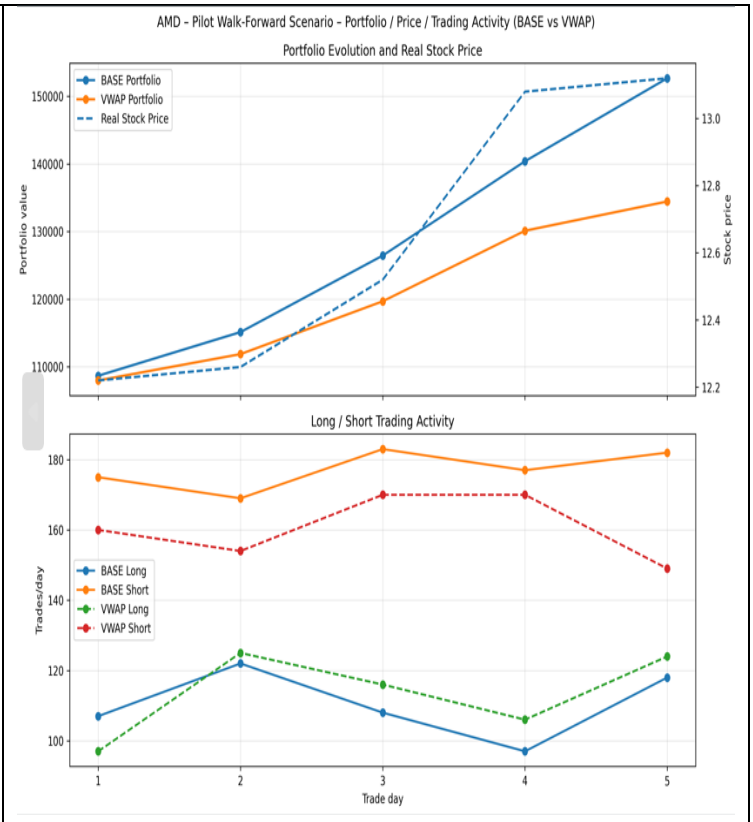
• στην πραγματική πορεία της τιμής της μετοχής,

• στην ημερήσια συναλλακτική δραστηριότητα long και short θέσεων.

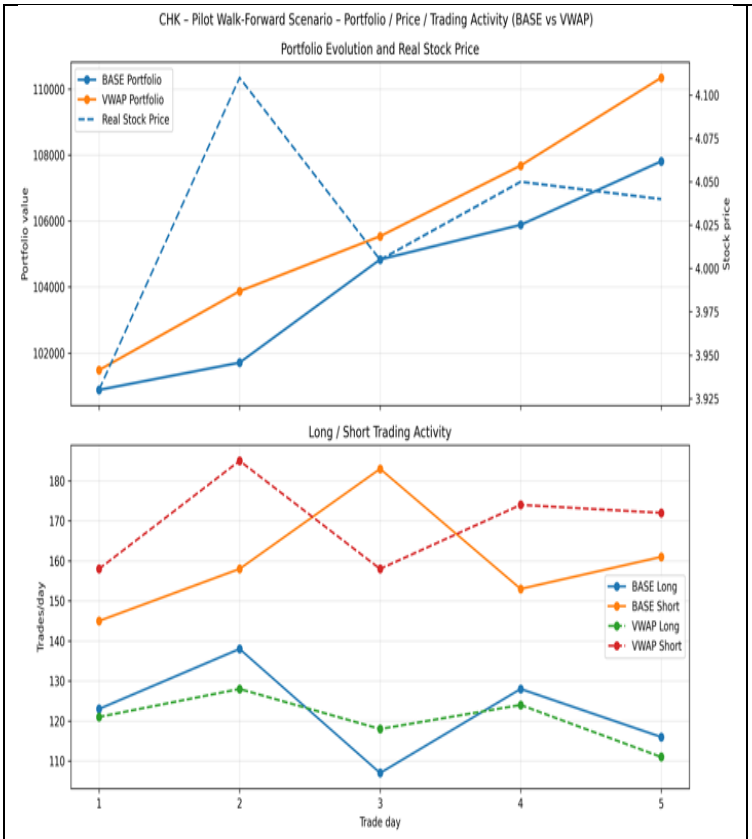
Για λόγους αναγνωσιμότητας και συγκριτικής παρουσίασης, τα αποτελέσματα αποτυπώνονται σε ημερήσια συγκεντρωτική μορφή, παρότι ο πράκτορας λαμβάνει αποφάσεις συναλλαγών σε κάθε χρονικό βήμα ενός λεπτού κατά τη διάρκεια κάθε χρηματιστηριακής συνεδρίασης.



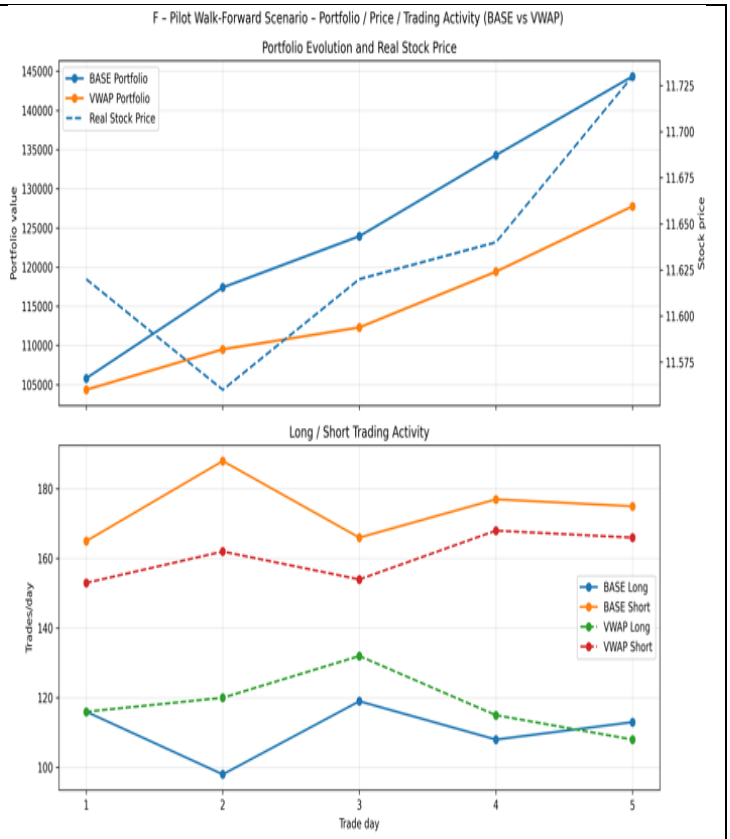
Σχήμα 5 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή AES στο Pilot Walk-Forward Scenario (BASE έναντι VWAP)



Σχήμα 6 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή AMD στο Pilot Walk-Forward Scenario (BASE έναντι VWAP)



Σχήμα 7 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή CHK στο Pilot Walk-Forward Scenario (BASE έναντι VWAP)



Σχήμα 8 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή F στο Pilot Walk-Forward Scenario (BASE έναντι VWAP)

Η συνδυαστική παρουσίαση χαρτοφυλακίου, πραγματικής τιμής αγοράς και συναλλακτικής δραστηριότητας επιτρέπει ουσιαστικότερη ερμηνεία της στρατηγικής συμπεριφοράς του πράκτορα. Η ανάλυση αυτή επιτρέπει τη σύνδεση της απόδοσης του χαρτοφυλακίου με τη συμπεριφορά της αγοράς και τη στρατηγική επιλογής ενεργειών, προσφέροντας πιο ολοκληρωμένη εικόνα της λειτουργίας του πράκτορα πέρα από απλές μετρικές απόδοσης. Από τα διαγράμματα προκύπτει ότι και στις δύο εκδοχές το χαρτοφυλάκιο εξελίσσεται δυναμικά κατά τη διάρκεια του μοναδικού πενθήμερου επεισοδίου, ξεκινώντας από κοινή αρχική αξία 100.000 δολαρίων. Η εξέλιξη αυτή αντανακλά τις αποφάσεις που λαμβάνει ο πράκτορας σε κάθε διαδοχική συνεδρία και δείχνει ότι είναι σε θέση να αναπτύσσει λειτουργικές στρατηγικές συναλλαγών στο συγκεκριμένο πειραματικό περιβάλλον.

Παράλληλα, η σύγκριση με την πραγματική πορεία της τιμής της μετοχής επιτρέπει την αξιολόγηση του κατά πόσο η στρατηγική ακολουθεί ή διαφοροποιείται από την υποκείμενη

αγοραία συμπεριφορά. Σε αρκετές περιπτώσεις, η εκδοχή BASE εμφανίζει πιο σταθερή και συνεπή ανοδική πορεία χαρτοφυλακίου, γεγονός που συνάδει με τις υψηλότερες τελικές αποδόσεις που καταγράφηκαν στον Πίνακα 5.1.

Η ανάλυση της ημερήσιας συναλλακτικής δραστηριότητας αναδεικνύει επιπλέον διαφοροποιήσεις στη στρατηγική λήψης αποφάσεων μεταξύ των δύο εκδοχών του συστήματος. Η βασική εκδοχή εμφανίζει συχνά μεγαλύτερη ένταση συναλλαγών, γεγονός που ενδέχεται να υποδηλώνει πιο επιθετική αξιοποίηση βραχυχρόνιων διακυμάνσεων της αγοράς. Αντίστοιχα, η VWAP εκδοχή παρουσιάζει σε αρκετές περιπτώσεις διαφορετική κατανομή long και short θέσεων, καθώς και τάση για πιο επιλεκτική ενεργοποίηση συναλλαγών. Η συμπεριφορά αυτή ενδέχεται να σχετίζεται με την πρόσθετη πληροφορία που παρέχουν τα χαρακτηριστικά VWAP σχετικά με τη σχέση τιμής και όγκου συναλλαγών, λειτουργώντας ως συμπληρωματικό σημείο αναφοράς κατά τη διαδικασία λήψης αποφάσεων. Παρότι η διαφοροποίηση αυτή δεν συνοδεύεται από συστηματική υπεροχή ως προς το τελικό PnL, θα μπορούσε να υποδηλώνει διαφορετικό προφίλ διαχείρισης συναλλαγών, το οποίο ενδέχεται να αποκτήσει μεγαλύτερη σημασία σε περιβάλλοντα όπου λαμβάνονται υπόψη πρόσθετοι παράγοντες, όπως το βάθος της αγοράς. Οι παρατηρούμενες διαφοροποιήσεις αφορούν τη συμπεριφορά της μαθημένης πολιτικής και όχι τον μηχανισμό εκτέλεσης συναλλαγών, ο οποίος παραμένει κοινός και στις δύο εκδοχές.

Συνεπώς, οι διαφορές απόδοσης μεταξύ BASE και VWAP αποδίδονται κυρίως:

• στην ποιότητα της αναπαράστασης κατάστασης,

• στη διαφοροποίηση της στρατηγικής επιλογής ενεργειών,

• στην προσαρμοστικότητα του πράκτορα στα δεδομένα αγοράς.

Συνολικά, το Pilot Walk-Forward Scenario (6-Day Dataset) επιβεβαιώνει ότι η προσθήκη χαρακτηριστικών VWAP δεν οδηγεί συστηματικά σε ανώτερη απόδοση, αλλά λειτουργεί ως εναλλακτικός μηχανισμός εμπλουτισμού της πληροφορίας εισόδου, του οποίου η αποτελεσματικότητα εξαρτάται από τη φύση της εκάστοτε μετοχής και τις συνθήκες της αγοράς.

Παράλληλα, το συγκεκριμένο σενάριο επιβεβαιώνει την ορθή λειτουργία της προτεινόμενης αρχιτεκτονικής και αποτελεί το πιλοτικό στάδιο της πειραματικής διαδικασίας, προετοιμάζοντας τη μετάβαση στο Extended Walk-Forward Scenario (111-Day Dataset), όπου

η ίδια μεθοδολογία walk-forward εφαρμόζεται σε σημαντικά μεγαλύτερο αριθμό ανεξάρτητων πενθήμερων επεισοδίων που καλύπτουν συνολικά 111 χρηματιστηριακές συνεδριάσεις και αποτελούν τη βάση για την εξαγωγή των τελικών αποτελεσμάτων της εργασίας.

5.2 Αποτελέσματα του Extended Walk-Forward Scenario (111-Day Dataset)

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα του δεύτερου και κύριου πειραματικού σεναρίου της εργασίας, το οποίο βασίζεται στο πλήρες σύνολο των 111 χρηματιστηριακών συνεδριάσεων (Extended Walk-Forward Scenario). Το συγκεκριμένο σενάριο αποτελεί το βασικό πλαίσιο αξιολόγησης του προτεινόμενου συστήματος ενισχυτικής μάθησης, καθώς εφαρμόζει την ίδια λογική walk-forward εκτέλεσης που παρουσιάστηκε στο πιλοτικό σενάριο, σε σημαντικά μεγαλύτερη χρονική κλίμακα και με πολύ μεγαλύτερο όγκο δεδομένων.

Όπως και στο πρώτο σενάριο, εξετάζονται δύο εκδοχές του συστήματος:

• η βασική εκδοχή (BASE),

• η εκδοχή με ενσωμάτωση πρόσθετων χαρακτηριστικών VWAP στο state vector (VWAP).

Η συνολική αρχιτεκτονική του Dueling Deep Q-Network, το Trading Environment, η Reward Function και οι μηχανισμοί εκπαίδευσης και βελτιστοποίησης παραμένουν απολύτως κοινοί, ώστε η συγκριτική αξιολόγηση να εστιάζει αποκλειστικά στην επίδραση της διαφοροποιημένης αναπαράστασης της κατάστασης.

Η απόδοση κάθε εκδοχής αποτιμάται μέσω της τελικής αξίας του χαρτοφυλακίου που προκύπτει από την εκτέλεση του πειραματικού σεναρίου, όπως καταγράφεται στα παραγόμενα αποτελέσματα του συστήματος

Μετοχή	BASE	VWAP
CHK	113.111,66	113.351,85
AMD	186.168,52	170.420,55
AES	175.888,26	173.199,69
F	180.521,78	174.453,18

Πίνακας 2 Τελική αξία χαρτοφυλακίου στο Extended Walk-Forward Scenario (111-Day Dataset)

Τα αποτελέσματα εμφανίζουν γενική συνέπεια με την εικόνα που προέκυψε και από το πρώτο πειραματικό σενάριο. Η ενσωμάτωση VWAP χαρακτηριστικών δεν οδηγεί σε συστηματικά ανώτερη απόδοση για όλες τις μετοχές. Στην περίπτωση της CHK καταγράφεται οριακά υψηλότερη τελική αξία υπέρ της VWAP εκδοχής, ενώ στις υπόλοιπες μετοχές η βασική εκδοχή επιτυγχάνει υψηλότερα τελικά επίπεδα χαρτοφυλακίου.

Η εικόνα αυτή υποδηλώνει ότι η προσθήκη VWAP λειτουργεί ως εναλλακτικός μηχανισμός εμπλουτισμού πληροφορίας και όχι ως καθολικός παράγοντας βελτίωσης της στρατηγικής. Η αποτελεσματικότητά του εξαρτάται από τη συμπεριφορά της εκάστοτε μετοχής, τη χρονική περίοδο και τη δυναμική των δεδομένων αγοράς.

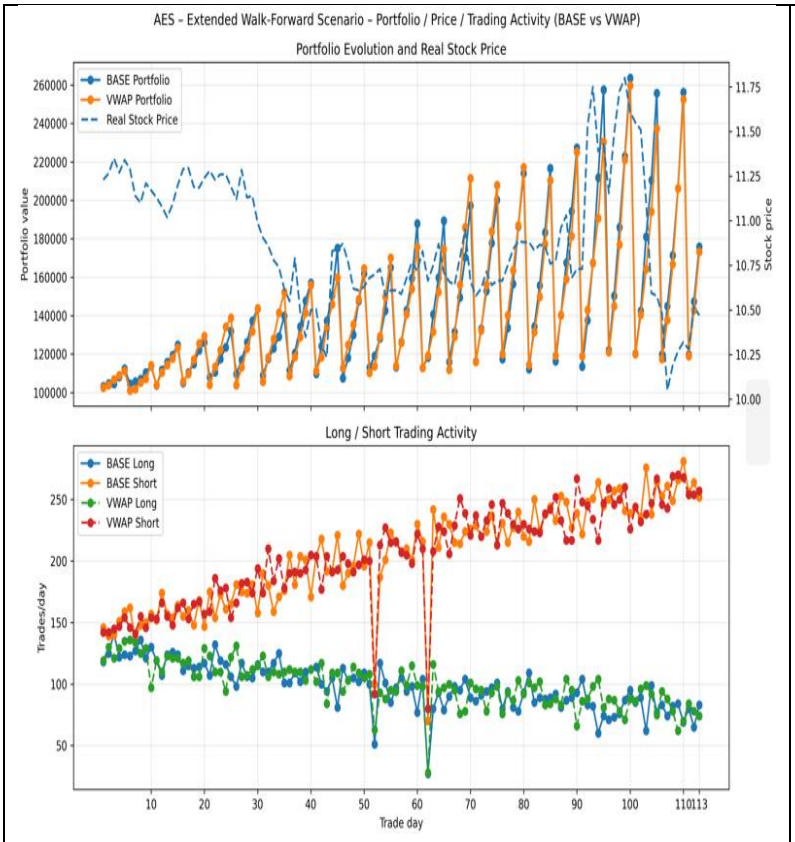
Για πληρέστερη αξιολόγηση, τα αποτελέσματα αναλύονται περαιτέρω μέσω:

• της εξέλιξης της αξίας του χαρτοφυλακίου,

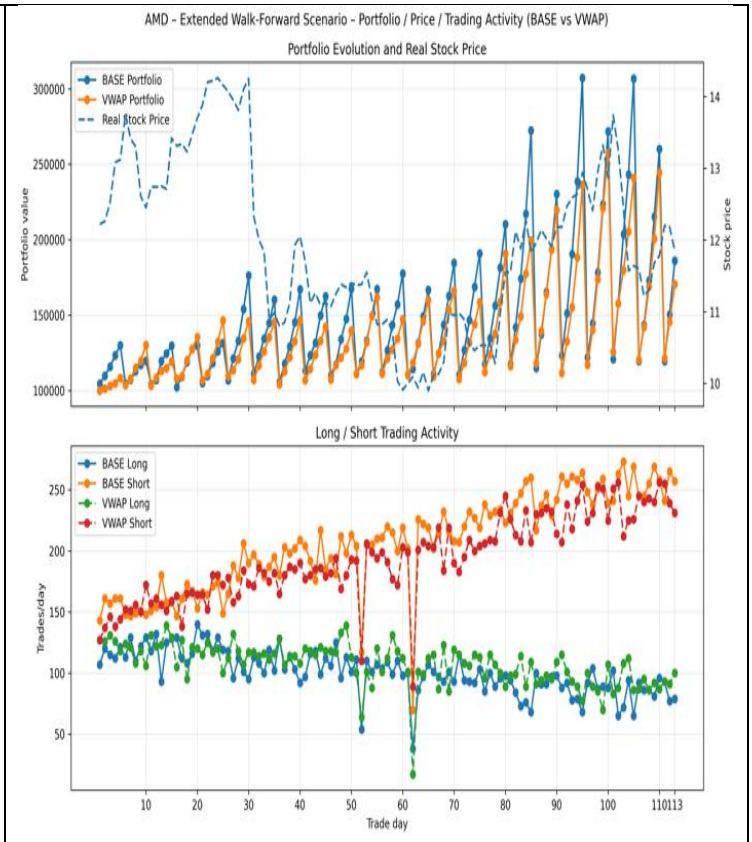
• της πραγματικής πορείας της τιμής κάθε μετοχής,

• της ημερήσιας δραστηριότητας long και short συναλλαγών.

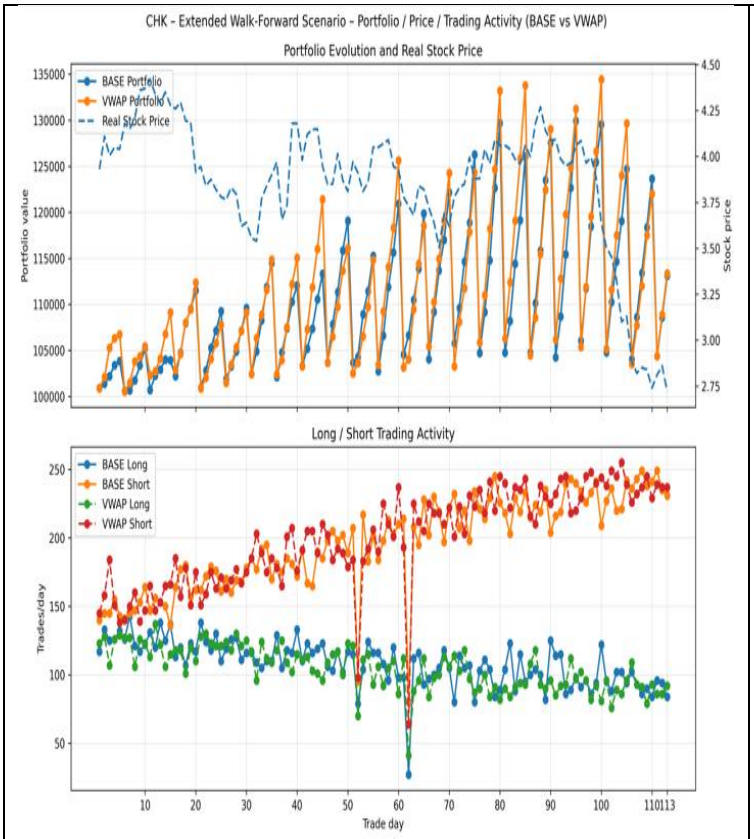
Για λόγους αναγνωσιμότητας και συγκριτικής παρουσίασης, τα αποτελέσματα αποτυπώνονται σε ημερήσια συγκεντρωτική μορφή, παρότι ο πράκτορας λαμβάνει αποφάσεις συναλλαγών σε κάθε χρονικό βήμα ενός λεπτού κατά τη διάρκεια κάθε χρηματιστηριακής συνεδρίασης.



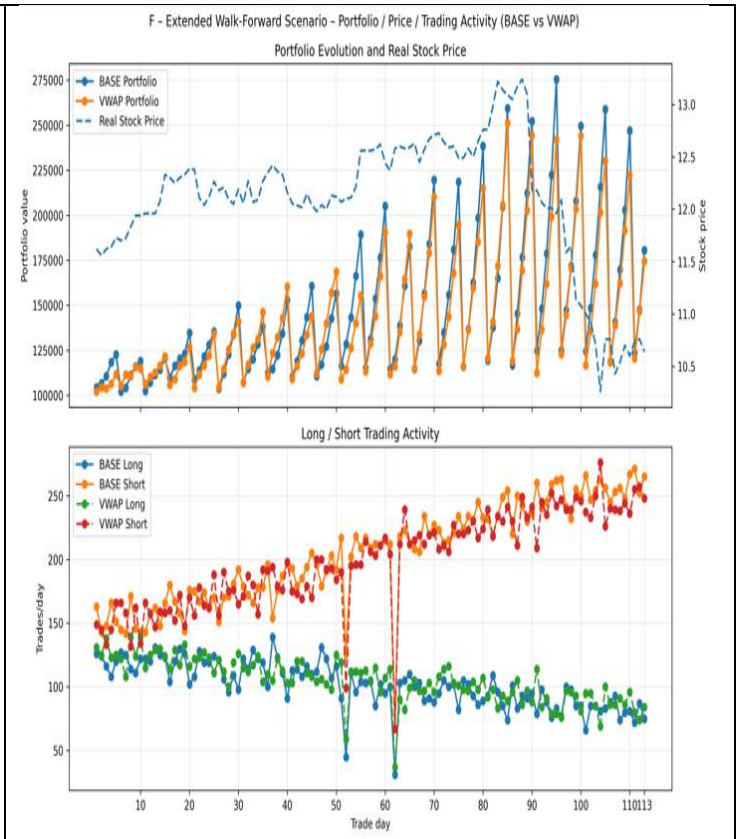
Σχήμα 9 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή AES στο Extended Walk-Forward Scenario (BASE έναντι VWAP)



Σχήμα 10 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή AMD στο Extended Walk-Forward Scenario (BASE έναντι VWAP)



Σχήμα 11 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή CHK στο Extended Walk-Forward Scenario (BASE έναντι VWAP)



Σχήμα 12 Εξέλιξη χαρτοφυλακίου, πραγματικής τιμής μετοχής και ημερήσιας συναλλακτικής δραστηριότητας (long/short θέσεις) για τη μετοχή F στο Extended Walk-Forward Scenario (BASE έναντι VWAP)

Από τα διαγράμματα εξέλιξης του χαρτοφυλακίου παρατηρούνται επαναλαμβανόμενοι κύκλοι ανόδου και επανεκκίνησης της αξίας του χαρτοφυλακίου. Η μορφή αυτή οφείλεται στη δομή του πειραματικού σεναρίου, σύμφωνα με την οποία κάθε επεισόδιο αποτελεί ανεξάρτητη πειραματική εκτέλεση. Στην αρχή κάθε επεισοδίου το περιβάλλον επανεκκινείται, το χαρτοφυλάκιο αρχικοποιείται εκ νέου με αρχικό κεφάλαιο 100.000 και ακολουθούν πέντε διαδοχικοί κύκλοι Train → Trade. Με την ολοκλήρωση του επεισοδίου η διαδικασία επαναλαμβάνεται από μεταγενέστερο χρονικό σημείο του dataset, χωρίς μεταφορά της αξίας του χαρτοφυλακίου από το προηγούμενο επεισόδιο.

Η ενσωμάτωση της πραγματικής τιμής της μετοχής στα διαγράμματα επιτρέπει ουσιαστικότερη σύγκριση μεταξύ της στρατηγικής του πράκτορα και της υποκείμενης συμπεριφοράς της αγοράς. Από τη σύγκριση αυτή προκύπτει ότι η εξέλιξη της αξίας του

χαρτοφυλακίου δεν αποτελεί απλή αντανάκλαση της πορείας της μετοχής, αλλά αποτέλεσμα της δυναμικής στρατηγικής long και short συναλλαγών που εφαρμόζει ο πράκτορας.

Η σύγκριση των εκδοχών BASE και VWAP αναδεικνύει ότι, παρά τις σημαντικές ομοιότητες στη συνολική στρατηγική συμπεριφορά, παρατηρούνται διαφοροποιήσεις στον τρόπο με τον οποίο ο πράκτορας αξιοποιεί την πληροφορία της αγοράς. Η βασική εκδοχή εμφανίζει συχνά μεγαλύτερη συναλλακτική δραστηριότητα, γεγονός που ενδέχεται να υποδηλώνει πιο επιθετική αξιοποίηση βραχυχρόνιων διακυμάνσεων της αγοράς. Αντίθετα, η εκδοχή VWAP παρουσιάζει σε αρκετές περιπτώσεις διαφορετική κατανομή long και short θέσεων, χωρίς όμως η διαφοροποίηση αυτή να μεταφράζεται συστηματικά σε υψηλότερη τελική απόδοση. Παράλληλα, τα διαγράμματα συναλλαγών δείχνουν ότι ο πράκτορας αξιοποιεί ενεργά τόσο ανοδικές όσο και πτωτικές κινήσεις της αγοράς, αναπτύσσοντας δυναμικές στρατηγικές διαπραγμάτευσης και όχι απλή παρακολούθηση της τάσης της μετοχής. Το εύρημα αυτό ενισχύει την άποψη ότι τα χαρακτηριστικά VWAP επηρεάζουν τη διαδικασία λήψης αποφάσεων του πράκτορα, χωρίς ωστόσο να αποτελούν από μόνα τους παράγοντα καθολικής βελτίωσης της απόδοσης.

Συνολικά, το Extended Walk-Forward Scenario (111-Day Dataset) παρέχει πληρέστερη αξιολόγηση του συστήματος σε σύγκριση με το πιλοτικό σενάριο, καθώς βασίζεται σε σημαντικά μεγαλύτερο χρονικό εύρος δεδομένων και σε μεγαλύτερο αριθμό κύκλων εκπαίδευσης και συναλλαγών. Παρά τη σημαντική αύξηση της κλίμακας του πειράματος, τα αποτελέσματα συγκλίνουν στο ίδιο βασικό συμπέρασμα: η προσθήκη χαρακτηριστικών VWAP δεν οδηγεί σε σταθερή ή καθολική βελτίωση της απόδοσης του πράκτορα.

Η βασική εκδοχή του συστήματος εμφανίζει γενικά συγκρίσιμη ή και υψηλότερη απόδοση στις περισσότερες περιπτώσεις, γεγονός που υποδηλώνει ότι το αρχικό σύνολο χαρακτηριστικών παρέχει ήδη σημαντική πληροφορία για τη μαθησιακή διαδικασία. Ωστόσο, οι διαφοροποιήσεις που παρατηρούνται στη συναλλακτική συμπεριφορά μεταξύ των δύο εκδοχών δείχνουν ότι η ενσωμάτωση των χαρακτηριστικών VWAP επηρεάζει τον τρόπο με τον οποίο ο πράκτορας διαμορφώνει τη στρατηγική του, ακόμη και όταν η επίδραση αυτή δεν αποτυπώνεται πάντοτε στην τελική αξία του χαρτοφυλακίου.

Συμπερασματικά, το δεύτερο πειραματικό σενάριο αποτελεί το κύριο πειραματικό πλαίσιο της εργασίας και ενισχύει ουσιαστικά την αξιοπιστία των συνολικών ευρημάτων. Η εφαρμογή της ίδιας μεθοδολογίας walk-forward σε διαδοχικά ανεξάρτητα επεισόδια που καλύπτουν

συνολικά 111 χρηματιστηριακές συνεδριάσεις επιτρέπει περισσότερο αντιπροσωπευτική αξιολόγηση της συμπεριφοράς του πράκτορα και παρέχει το βασικό υπόβαθρο για την τελική συγκριτική αποτίμηση της συμβολής του δείκτη VWAP στη λειτουργία του Dueling Deep Q-Network.

5.3Συγκριτική ανάλυση BASE και VWAP

Η συγκριτική αξιολόγηση των δύο πειραματικών σεναρίων της παρούσας εργασίας αναδεικνύει ότι η ενσωμάτωση πρόσθετων χαρακτηριστικών VWAP στο διάνυσμα κατάστασης του πράκτορα δεν οδηγεί σε συστηματική και καθολική βελτίωση της απόδοσης του συστήματος ενισχυτικής μάθησης. Παρότι σε ορισμένες επιμέρους περιπτώσεις παρατηρούνται βελτιωμένα αποτελέσματα υπέρ της VWAP εκδοχής, η συνολική εικόνα δείχνει ότι η BASE εκδοχή εμφανίζει συχνά συγκρίσιμη ή και ανώτερη απόδοση σε αρκετές από τις εξεταζόμενες περιπτώσεις, χωρίς όμως η υπεροχή αυτή να εμφανίζεται καθολικά σε όλα τα πειραματικά σενάρια και τις μετοχές.

Η παρατήρηση αυτή εμφανίζεται με συνέπεια τόσο στο Pilot Walk-Forward Scenario (6-Day Dataset) όσο και στο Extended Walk-Forward Scenario (111-Day Dataset), γεγονός που ενισχύει την αξιοπιστία του συμπεράσματος. Η επαναληψιμότητα των αποτελεσμάτων σε δύο πειραματικά σενάρια διαφορετικής κλίμακας υποδηλώνει ότι η επίδραση του VWAP δεν περιορίζεται στα χαρακτηριστικά ενός συγκεκριμένου συνόλου δεδομένων, αλλά σχετίζεται ευρύτερα με τη φύση της αναπαράστασης των χαρακτηριστικών και της λειτουργίας του συστήματος.

Παράλληλα, τα αποτελέσματα δείχνουν ότι η επίδραση του VWAP διαφοροποιείται μεταξύ διαφορετικών μετοχών και συνθηκών αγοράς. Σε ορισμένες περιπτώσεις, όπως σε επιμέρους αξιολογήσεις της AMD και της CHK, η εκδοχή VWAP παρουσίασε ανταγωνιστική ή και βελτιωμένη συμπεριφορά σε σχέση με την BASE εκδοχή, ενώ σε άλλες περιπτώσεις η BASE εκδοχή εμφάνισε υψηλότερη συνολική κερδοφορία. Η παρατήρηση αυτή υποδηλώνει ότι η αποτελεσματικότητα του VWAP δεν μπορεί να θεωρηθεί γενικεύσιμη σε όλες τις περιπτώσεις, αλλά εξαρτάται από τα ιδιαίτερα χαρακτηριστικά κάθε χρηματοοικονομικού περιβάλλοντος, όπως:

- η μεταβλητότητα της αγοράς,
- η δομή του όγκου συναλλαγών,
- η ένταση των ανοδικών ή καθοδικών τάσεων,
- και η δυναμική της σχέσης τιμής–όγκου.

Η ανάλυση των αποτελεσμάτων δείχνει επίσης ότι οι δύο εκδοχές του πράκτορα παρουσιάζουν διαφοροποιημένη συμπεριφορά ως προς τη στρατηγική συναλλαγών. Σε αρκετές περιπτώσεις,

η εκδοχή VWAP πραγματοποίησε μικρότερο αριθμό συναλλαγών σε σχέση με την BASE εκδοχή, στοιχείο που ενδέχεται να συνδέεται με πιο επιφυλακτική ή συντηρητική στρατηγική λήψης αποφάσεων. Αντίθετα, η BASE εκδοχή εμφάνισε συχνότερα πιο επιθετική συμπεριφορά, εκμεταλλευόμενη εντονότερα βραχυχρόνιες μεταβολές της αγοράς με στόχο τη μεγιστοποίηση της συνολικής κερδοφορίας.

Η διαφοροποίηση αυτή εμφανίζεται εντονότερα σε περιόδους αυξημένης μεταβλητότητας ή απότομων πτωτικών κινήσεων της αγοράς. Ειδικότερα, στο Extended Walk-Forward Scenario (111-Day Dataset), η εκδοχή VWAP παρουσίασε σε ορισμένες περιπτώσεις διαφοροποιημένη συναλλακτική συμπεριφορά κατά τη διάρκεια έντονων καθοδικών μεταβολών της αγοράς, με τάση για περιορισμένη δραστηριότητα σε σχέση με την BASE εκδοχή. Το εύρημα αυτό υποδηλώνει ότι η ενσωμάτωση πληροφορίας όγκου μέσω του VWAP ενδέχεται να επηρεάζει τη συμπεριφορά του πράκτορα προς πιο συντηρητικές αποφάσεις σε συνθήκες αυξημένης αβεβαιότητας, χωρίς ωστόσο να οδηγεί απαραίτητα σε υψηλότερο συνολικό Profit and Loss (PnL).

Από την άλλη πλευρά, τα αποτελέσματα δείχνουν επίσης ότι σε έντονα ανοδικές περιόδους αγοράς η BASE εκδοχή συχνά επιτυγχάνει υψηλότερη συνολική κερδοφορία, πιθανώς λόγω μεγαλύτερης επιθετικότητας στην εκμετάλλευση βραχυχρόνιων τάσεων. Το γεγονός αυτό αναδεικνύει ότι η επίδραση του VWAP δεν μπορεί να αξιολογηθεί αποκλειστικά με όρους συνολικού PnL, αλλά απαιτεί εξέταση της συνολικής συμπεριφοράς του πράκτορα και της στρατηγικής λήψης αποφάσεων που προκύπτει από τη μαθησιακή διαδικασία.

Μία πιθανή ερμηνεία των αποτελεσμάτων είναι ότι ο δείκτης VWAP προκύπτει από συνδυασμό της πληροφορίας της τιμής και του όγκου συναλλαγών, στοιχεία τα οποία περιλαμβάνονται ήδη στα πρωτογενή δεδομένα εισόδου του συστήματος. Επομένως, είναι πιθανό το νευρωνικό δίκτυο να μπορεί, κατά τη διαδικασία εκπαίδευσης, να εξάγει έμμεσα παρόμοια πληροφορία από τα διαθέσιμα χαρακτηριστικά, χωρίς να απαιτείται η ρητή ενσωμάτωση του VWAP ως πρόσθετου χαρακτηριστικού στο state vector.

Κατά συνέπεια, η προσθήκη του VWAP ενδέχεται να εισάγει πληροφορία που παρουσιάζει σημαντική συσχέτιση με τα ήδη διαθέσιμα χαρακτηριστικά. Στην περίπτωση αυτή, η αύξηση της διάστασης του χώρου καταστάσεων δεν συνεπάγεται κατ’ ανάγκη αντίστοιχη αύξηση της πραγματικής πληροφοριακής αξίας που είναι διαθέσιμη στον πράκτορα, ενώ είναι πιθανό να

αυξάνει την πολυπλοκότητα της μαθησιακής διαδικασίας χωρίς ανάλογη βελτίωση της απόδοσης.

Η αύξηση της διάστασης του χώρου καταστάσεων αποτελεί επίσης κρίσιμο τεχνικό παράγοντα. Η ενσωμάτωση πρόσθετων χαρακτηριστικών αυξάνει τη συνθετότητα της μαθησιακής αναπαράστασης, απαιτώντας από το μοντέλο να προσαρμοστεί σε μεγαλύτερο χώρο εισόδου. Σε περιβάλλοντα όπου τα διαθέσιμα δεδομένα ή οι περίοδοι εκπαίδευσης παραμένουν πεπερασμένα, η αύξηση αυτή ενδέχεται να περιορίσει την αποτελεσματικότητα της μάθησης αντί να την ενισχύσει.

Επιπλέον, η συγκεκριμένη υλοποίηση του trading environment διαφοροποιείται ουσιαστικά από πραγματικά execution frameworks στα οποία ο VWAP χρησιμοποιείται παραδοσιακά ως δείκτης ποιότητας εκτέλεσης. Στην παρούσα εργασία, ο VWAP δεν χρησιμοποιείται ως κριτήριο εκτέλεσης συναλλαγών ή ως benchmark ποιότητας εκτέλεσης, αλλά αποκλειστικά ως πρόσθετο πληροφοριακό χαρακτηριστικό εισόδου. Οι συναλλαγές του πράκτορα εκτελούνται βάσει της υφιστάμενης λογικής του περιβάλλοντος και όχι βάσει βελτιστοποίησης εκτέλεσης ως προς VWAP benchmarks.

Κατά συνέπεια, η ενσωμάτωση VWAP δεν αξιοποιεί πλήρως τον κλασικό επιχειρησιακό ρόλο του δείκτη στις πραγματικές αγορές, όπου χρησιμοποιείται συχνά για:

- μέτρηση ποιότητας εκτέλεσης,
- αξιολόγηση κόστους συναλλαγών,
- εκτίμηση ρευστότητας,
- παρακολούθηση intraday price-volume dynamics.

Η απουσία δεδομένων μικροδομής αγοράς, όπως bid prices, ask prices και order book depth περιορίζει περαιτέρω τη δυνατότητα πλήρους αξιοποίησης του VWAP ως στρατηγικού πλεονεκτήματος.

Συνολικά, η σύγκριση μεταξύ BASE και VWAP αναδεικνύει ένα ουσιαστικότερο επιστημονικό συμπέρασμα: η προσθήκη επιπλέον χαρακτηριστικών εισόδου σε ένα σύστημα βαθιάς ενισχυτικής μάθησης δεν εγγυάται αυτομάτως βελτιωμένη απόδοση.

Η πραγματική αξία κάθε νέου χαρακτηριστικού εξαρτάται από:

• τη συμπληρωματικότητα της πληροφορίας,

• τη συσχέτισή του με υπάρχοντα δεδομένα,

• τη συμβατότητα με τη δομή του περιβάλλοντος,

• τη φύση της reward function,

• και τη δυνατότητα ουσιαστικής αξιοποίησής του από τη μαθησιακή αρχιτεκτονική.

Τα ευρήματα της παρούσας εργασίας υπογραμμίζουν τη σημασία της προσεκτικής επιλογής χαρακτηριστικών, της εμπειρικής αξιολόγησης και της ισορροπίας μεταξύ πολυπλοκότητας και πληροφοριακής αξίας κατά τον σχεδιασμό προηγμένων συστημάτων αλγοριθμικής διαπραγμάτευσης.

Συμπερασματικά, τα αποτελέσματα των δύο πειραματικών σεναρίων συγκλίνουν στο ίδιο βασικό εύρημα. Η βασική εκδοχή του συστήματος παρέχει ήδη μία ισχυρή πληροφοριακή βάση για τη μαθησιακή διαδικασία, ενώ η ενσωμάτωση των χαρακτηριστικών VWAP λειτουργεί κυρίως ως μηχανισμός εμπλουτισμού της αναπαράστασης της κατάστασης και διαφοροποίησης της στρατηγικής συμπεριφοράς του πράκτορα, χωρίς όμως να οδηγεί συστηματικά σε υψηλότερη τελική απόδοση.

Το αποτέλεσμα αυτό αναδεικνύει μία ευρύτερη αρχή στον σχεδιασμό συστημάτων βαθιάς ενισχυτικής μάθησης για χρηματοοικονομικές εφαρμογές: η αποτελεσματικότητα ενός νέου χαρακτηριστικού δεν εξαρτάται μόνο από τη χρηματοοικονομική του σημασία, αλλά και από τον βαθμό στον οποίο προσφέρει νέα, συμπληρωματική πληροφορία σε σχέση με τα ήδη διαθέσιμα χαρακτηριστικά και μπορεί να αξιοποιηθεί αποτελεσματικά από τη μαθησιακή αρχιτεκτονική.

5.4 Συμπεράσματα

Στο παρόν κεφάλαιο παρουσιάστηκαν και αναλύθηκαν τα αποτελέσματα των δύο βασικών πειραματικών σεναρίων της εργασίας, με στόχο τόσο την αποτίμηση της συνολικής αποτελεσματικότητας του προτεινόμενου συστήματος ενισχυτικής μάθησης όσο και τη διερεύνηση της επίδρασης της ενσωμάτωσης πρόσθετων χαρακτηριστικών VWAP στο state representation του πράκτορα.

Η ανάλυση των αποτελεσμάτων επιβεβαίωσε ότι ο πράκτορας Dueling Deep Q-Network είναι σε θέση να αναπτύσσει λειτουργικές στρατηγικές αλγοριθμικής διαπραγμάτευσης, αξιοποιώντας ιστορικά δεδομένα αγοράς και προσαρμόζοντας σταδιακά τη συμπεριφορά του μέσω της αλληλεπίδρασης με το περιβάλλον συναλλαγών. Και στα δύο πειραματικά σενάρια, ο πράκτορας παρουσίασε ικανότητα παραγωγής θετικής εξέλιξης χαρτοφυλακίου, γεγονός που καταδεικνύει την επιχειρησιακή βιωσιμότητα της προτεινόμενης αρχιτεκτονικής.

Το πρώτο πειραματικό σενάριο, Pilot Walk-Forward Scenario (6-Day Dataset), λειτούργησε λειτούργησε ως πιλοτικό πλαίσιο αρχικής αξιολόγησης, επιτρέποντας την επαλήθευση της λειτουργικότητας του προτεινόμενου συστήματος και την πρώτη συγκριτική αξιολόγηση μεταξύ των εκδοχών BASE και VWAP σε περιορισμένο αλλά πλήρως ελεγχόμενο σύνολο δεδομένων. Τα αποτελέσματα του σεναρίου ανέδειξαν ότι τόσο η BASE όσο και η VWAP εκδοχή μπορούν να επιτύχουν θετικές αποδόσεις, χωρίς όμως να προκύπτει σαφής και συστηματική υπεροχή κάποιας από τις δύο προσεγγίσεις. Η διαφοροποίηση των αποτελεσμάτων μεταξύ των μετοχών υποδηλώνει ότι η συμβολή του VWAP εξαρτάται από τα ιδιαίτερα χαρακτηριστικά κάθε χρηματοοικονομικού περιβάλλοντος.

Το δεύτερο και κύριο πειραματικό σενάριο, Extended Walk-Forward Scenario (111-Day Dataset), αποτέλεσε το βασικό πλαίσιο αξιολόγησης της εργασίας. Εφαρμόζοντας την ίδια μεθοδολογία walk-forward σε διαδοχικά ανεξάρτητα επεισόδια που καλύπτουν συνολικά 111 χρηματιστηριακές συνεδριάσεις, παρείχε περισσότερο αντιπροσωπευτική εικόνα της συμπεριφοράς του πράκτορα σε διαφορετικές συνθήκες της αγοράς. Η μεγαλύτερη κλίμακα του πειράματος ενίσχυσε την αξιοπιστία των αποτελεσμάτων και επιβεβαίωσε ότι τα βασικά συμπεράσματα που προέκυψαν από το πιλοτικό σενάριο διατηρούνται και στο εκτεταμένο πειραματικό πλαίσιο.

Η ανάλυση της συναλλακτικής συμπεριφοράς έδειξε ότι ο πράκτορας αναπτύσσει συνεκτικά πρότυπα λήψης αποφάσεων, με σαφή αξιοποίηση τόσο long όσο και short στρατηγικών. Η σχετική υπεροχή των short συναλλαγών σε πολλές περιπτώσεις υποδηλώνει προσαρμογή σε συγκεκριμένες δυναμικές αγοράς και δεν συνδέεται αποκλειστικά με την παρουσία του VWAP, αλλά με τη συνολική μαθημένη πολιτική του συστήματος.

Παράλληλα, παρατηρήθηκε ότι οι δύο εκδοχές του πράκτορα παρουσίασαν διαφοροποιημένη συμπεριφορά ως προς τον αριθμό και την κατανομή των συναλλαγών. Σε αρκετές περιπτώσεις, η εκδοχή VWAP πραγματοποίησε μικρότερο αριθμό συναλλαγών, στοιχείο που ενδέχεται να συνδέεται με πιο επιφυλακτική στρατηγική λήψης αποφάσεων, ιδιαίτερα σε περιόδους αυξημένης μεταβλητότητας. Αντίθετα, η BASE εκδοχή παρουσίασε συχνότερα πιο επιθετική αξιοποίηση βραχυχρόνιων τάσεων της αγοράς, γεγονός που σε αρκετές περιπτώσεις συνδέθηκε με υψηλότερη συνολική κερδοφορία.

Ένα από τα σημαντικότερα ευρήματα της εργασίας αφορά τη συγκριτική αξιολόγηση μεταξύ BASE και VWAP. Η συνολική εικόνα των αποτελεσμάτων καταδεικνύει ότι η προσθήκη VWAP χαρακτηριστικών δεν οδηγεί σε συστηματική ή καθολική βελτίωση της απόδοσης. Παρά τις επιμέρους διαφοροποιήσεις, η βασική εκδοχή του μοντέλου εμφανίζει συχνά συγκρίσιμη ή ανώτερη συμπεριφορά σε αρκετές περιπτώσεις, χωρίς όμως η υπεροχή αυτή να εμφανίζεται καθολικά σε όλα τα πειραματικά σενάρια και τις μετοχές.

Το αποτέλεσμα αυτό αποδίδεται πιθανότατα στους εξής βασικούς παράγοντες:

• ο VWAP παράγεται από πληροφορία ήδη διαθέσιμη στα δεδομένα τιμής και όγκου,

• η πληροφορία του μπορεί να είναι μερικώς πλεονάζουσα,

• η αύξηση της διάστασης του state vector ενδέχεται να αυξάνει την πολυπλοκότητα χωρίς αντίστοιχο πληροφοριακό όφελος,

• η παρούσα προσομοίωση δεν αξιοποιεί τον VWAP ως execution benchmark αλλά μόνο ως πληροφοριακό χαρακτηριστικό.

Συνεπώς, η εργασία αναδεικνύει ένα ιδιαίτερα σημαντικό ερευνητικό συμπέρασμα: η ενσωμάτωση νέων χαρακτηριστικών στο state vector ενός συστήματος βαθιάς ενισχυτικής μάθησης δεν συνεπάγεται αυτομάτως βελτίωση της απόδοσης. Η πραγματική συνεισφορά κάθε νέου χαρακτηριστικού πρέπει να αξιολογείται εμπειρικά, λαμβάνοντας υπόψη τόσο τη συμπληρωματικότητα της πληροφορίας που προσφέρει όσο και τον τρόπο με τον οποίο αυτή

αξιοποιείται από τη μαθησιακή αρχιτεκτονική. Η αποτελεσματικότητα κάθε νέου χαρακτηριστικού εξαρτάται ουσιαστικά από:

• τη συμπληρωματικότητα της πληροφορίας,

• τη συσχέτισή του με υπάρχοντα δεδομένα,

• τη δομή του περιβάλλοντος,

• την αρχιτεκτονική του μοντέλου,

• τη συμβατότητα με τη reward function.

Με βάση τα συνολικά ευρήματα, τα βασικά ερευνητικά ερωτήματα της εργασίας απαντώνται ως εξής:

• Η ενισχυτική μάθηση μπορεί να εφαρμοστεί αποτελεσματικά σε περιβάλλοντα αλγοριθμικής διαπραγμάτευσης.

• Ο πράκτορας Dueling DQN είναι ικανός να αναπτύσσει λειτουργικές και αποδοτικές στρατηγικές συναλλαγών.

• Η ενσωμάτωση VWAP ως πρόσθετου χαρακτηριστικού δεν παρέχει συστηματικά ανώτερη απόδοση στο συγκεκριμένο πλαίσιο.

• Η επιλογή και η αξιολόγηση των χαρακτηριστικών εισόδου αποτελεί κρίσιμο παράγοντα σχεδιασμού συστημάτων βαθιάς ενισχυτικής μάθησης.

Συμπερασματικά, η παρούσα εργασία καταδεικνύει ότι τα συστήματα βαθιάς ενισχυτικής μάθησης διαθέτουν σημαντικές δυνατότητες εφαρμογής στον χώρο της αλγοριθμικής διαπραγμάτευσης, υπό την προϋπόθεση προσεκτικού σχεδιασμού της αρχιτεκτονικής, του χώρου καταστάσεων και της πειραματικής διαδικασίας αξιολόγησης.

Τα ευρήματα ενισχύουν τη σημασία:

• της συστηματικής εμπειρικής αξιολόγησης,

• της προσεκτικής επιλογής χαρακτηριστικών,

• της ισορροπίας μεταξύ πολυπλοκότητας και πληροφοριακής αξίας,

• της εμπειρικής επικύρωσης σε πολλαπλά πειραματικά πλαίσια.

Με τον τρόπο αυτό, η εργασία συμβάλλει ουσιαστικά στη μελέτη εφαρμογής προηγμένων τεχνικών ενισχυτικής μάθησης σε χρηματοοικονομικά περιβάλλοντα και δημιουργεί σταθερή βάση για μελλοντική ερευνητική επέκταση.

5.5 Μελλοντικές επεκτάσεις

Τα αποτελέσματα της παρούσας εργασίας αναδεικνύουν μια σειρά από κατευθύνσεις για μελλοντική έρευνα και βελτίωση του προτεινόμενου συστήματος.

Μια πρώτη κατεύθυνση αφορά τον εμπλουτισμό του συνόλου χαρακτηριστικών που συνθέτουν το διάνυσμα κατάστασης (state vector) του πράκτορα. Η παρούσα προσέγγιση βασίζεται κυρίως σε δεδομένα τιμών και όγκου συναλλαγών, καθώς και σε περιορισμένο αριθμό τεχνικών δεικτών. Η ενσωμάτωση επιπλέον χαρακτηριστικών, όπως δεδομένα μικροδομής της αγοράς (π.χ. βάθος βιβλίου εντολών (order book depth) και ανισορροπία εντολών (order imbalance)), καθώς και μακροοικονομικά ή διακλαδικά χαρακτηριστικά, θα μπορούσε να εμπλουτίσει την αναπαράσταση της κατάστασης και να επιτρέψει στον πράκτορα να λαμβάνει αποφάσεις αξιοποιώντας πληρέστερη πληροφορία για το περιβάλλον της αγοράς.

Μια δεύτερη σημαντική κατεύθυνση αφορά την ανάπτυξη ενός περισσότερο ρεαλιστικού περιβάλλοντος προσομοίωσης των συναλλαγών (Trading Environment). Στην παρούσα υλοποίηση οι συναλλαγές εκτελούνται με απλοποιημένο τρόπο στο τέλος κάθε χρονικού βήματος, χωρίς να λαμβάνονται υπόψη παράγοντες που επηρεάζουν την πραγματική εκτέλεση των εντολών, όπως το bid–ask spread, η μερική εκτέλεση εντολών (partial fills), οι περιορισμοί ρευστότητας ή ο αντίκτυπος των συναλλαγών στην αγορά (market impact). Η ενσωμάτωση τέτοιων μηχανισμών θα επέτρεπε ακριβέστερη αποτίμηση της πραγματικής αποτελεσματικότητας των στρατηγικών και θα προσέγγιζε περισσότερο τις συνθήκες λειτουργίας πραγματικών συστημάτων αλγοριθμικής διαπραγμάτευσης.

Επιπλέον, ενδιαφέρον παρουσιάζει η διερεύνηση εναλλακτικών αρχιτεκτονικών ενισχυτικής μάθησης. Παρότι η αρχιτεκτονική Dueling Deep Q-Network απέδωσε ικανοποιητικά αποτελέσματα, η αξιολόγηση νεότερων προσεγγίσεων, όπως Double DQN, Rainbow DQN, Proximal Policy Optimization (PPO), Soft Actor–Critic (SAC) ή άλλες σύγχρονες Actor–Critic αρχιτεκτονικές, θα μπορούσε να προσφέρει πρόσθετα στοιχεία σχετικά με την καταλληλότητά τους για προβλήματα αλγοριθμικής διαπραγμάτευσης.

Τέλος, ιδιαίτερο ενδιαφέρον θα παρουσίαζε η επέκταση της προσέγγισης σε περιβάλλον πολλαπλών μετοχών. Στην παρούσα υλοποίηση κάθε μετοχή εξετάζεται ανεξάρτητα, χωρίς αλληλεπίδραση μεταξύ τους. Η ανάπτυξη ενός πράκτορα που διαχειρίζεται ταυτόχρονα ένα

χαρτοφυλάκιο πολλών τίτλων θα επέτρεπε τη μελέτη στρατηγικών κατανομής κεφαλαίου, διαφοροποίησης κινδύνου και δυναμικής ανακατανομής πόρων μεταξύ περιουσιακών στοιχείων. Μια τέτοια προσέγγιση θα επέτρεπε επίσης τη διερεύνηση της εκμάθησης στρατηγικών κατανομής κεφαλαίου σε επίπεδο χαρτοφυλακίου, αντί της ανεξάρτητης διαχείρισης κάθε μετοχής.

Συνολικά, οι παραπάνω κατευθύνσεις αναδεικνύουν ότι, παρότι η παρούσα εργασία παρέχει ένα λειτουργικό και μεθοδολογικά συνεκτικό πλαίσιο εφαρμογής της βαθιάς ενισχυτικής μάθησης σε ενδοημερήσια χρηματιστηριακά δεδομένα, εξακολουθούν να υπάρχουν σημαντικά περιθώρια επέκτασης και βελτίωσης. Η διερεύνηση πλουσιότερων συνόλων χαρακτηριστικών, η υιοθέτηση ρεαλιστικότερων συνθηκών προσομοίωσης, η αξιολόγηση εναλλακτικών αρχιτεκτονικών ενισχυτικής μάθησης και η επέκταση σε διαχείριση πολυμετοχικών χαρτοφυλακίων μπορούν να συμβάλουν στην ανάπτυξη πιο αποδοτικών, πιο ανθεκτικών και περισσότερο εφαρμόσιμων συστημάτων αλγοριθμικής διαπραγμάτευσης.

Βιβλιογραφία

Bialkowski, J., Darolles, S. and Le Fol, G. (2012) 'Reducing the Risk of VWAP Orders Execution: A New Approach to Modelling Intra-Day Volume', JASSA, 1.

Bouchard, B. and Dang, N.M. (2013) 'Generalized Stochastic Target Problems for Pricing and Partial Hedging under Loss Constraints: Application in Optimal Book Liquidation', Finance and Stochastics, 17(1), pp. 31–72.

Chen, R. (2024) 'A Review of VWAP Trading Algorithms: Development, Improvements and Limitations'. Proceedings of the 3rd International Conference on Financial Technology and Business Analysis, pp. 185–191. doi:10.54254/2754-1169/135/2024.18582.

Frei, C. and Westray, N. (2015) 'Optimal Execution of a VWAP Order: A Stochastic Control Approach', Mathematical Finance, 25(3), pp. 612–639.

Genet, R. (2025) 'Deep Learning for VWAP Execution in Crypto Markets: Beyond the Volume Curve', arXiv preprint arXiv:2502.13722.

Genet, R. (2025) 'Recurrent Neural Networks for Dynamic VWAP Execution: Adaptive Trading Strategies with Temporal Kolmogorov-Arnold Networks', arXiv preprint, arXiv:2502.18177.

Glavic, M. et al. (2017) 'Reinforcement learning for energy management in buildings: A survey', *Energy and Buildings*, 139, pp. 1–12.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning.* MIT Press.

Hasselt, H. V. (2010) 'Double Q-Learning', *Advances in Neural Information Processing Systems*, 23.

Hu, G. (2023) 'Advancing Algorithmic Trading: A Multi-Technique Enhancement of Deep Q-Network Models', arXiv preprint, arXiv:2311.05743.

Huang, Y., Zhou, C., Zhang, L. and Lu, X. (2024) 'A Self-Rewarding Mechanism in Deep Reinforcement Learning for Trading Strategy Optimization', Mathematics, 12(24), Article 4020. doi:10.3390/math12244020.

Humphery-Jenner, M.L. (2011) 'Optimal VWAP Trading under Noisy Conditions', Journal of Banking and Finance, 35(9), pp. 2319–2329.

Hyndman, R.J. (2011) *Cyclic and Seasonal Time Series.* Monash University.

Jinshimo, A., Liang, D., Lin, Z. and Wang, C. (2023) 'The Application of Deep Reinforcement Learning in Stock Trading Models', *Proceedings of the 7th International Conference on Economic Management and Green Development*, pp. 215–223. doi:10.54254/2754-1169/39/20231973.

Kaelbling, L., Littman, M. and Moore, A. (1996) 'Reinforcement Learning: A Survey', *Journal of Artificial Intelligence Research*, 4, pp. 237–285.

Kim, S., Kim, J., Sul, H.K. and Hong, Y. (2024) 'An adaptive dual-level reinforcement learning approach for optimal trade execution', Expert Systems with Applications, 252, 124263.

Li, S. (2023) *Reinforcement Learning for Sequential Decision and Optimal Control.* Springer.

Li, Y. (2018) *Deep Reinforcement Learning: An Overview.* arXiv:1812.05551.

Li, Z. et al. (2023) 'Stock Market Analysis and Prediction using LSTM', *Procedia Computer Science*, 218, pp. 421–430.

Liang, Z. and Chen, W. (2018) *Adversarial Deep Reinforcement Learning in Portfolio Management.* arXiv:1808.09940.

Mahmoodzadeh, Z. et al. (2020) 'Condition-Based Maintenance with Reinforcement Learning for Dry Gas Pipelines Subject to Internal Corrosion', *Sensors*, 20(7), pp. 2001–2019.

McCulloch, J. and Kazakov, V. (2012) Mean Variance Optimal VWAP Trading. Available at: https://ssrn.com/abstract=1803858.

Melo, F. (2001) *Convergence of Q-Learning: A Simple Proof.* Technical Report.

Mnih, V. et al. (2013) *Playing Atari with Deep Reinforcement Learning.* arXiv:1312.5602.

Mnih, V. et al. (2015) 'Human-level control through deep reinforcement learning', *Nature*, 518(7540), pp. 529–533.

Moody, J. and Saffell, M. (1998) 'Performance Functions and Reinforcement Learning for Trading Systems and Portfolios', *Advances in Neural Information Processing Systems*, 10, pp. 917–923.

Murphy, J.J. (1999) *Technical Analysis of the Financial Markets.* New York Institute of Finance.

Olorunnimbe, K. and Viktor, H. (2023) 'Deep learning in the stock market: A systematic survey of practice, backtesting, and applications', Artificial Intelligence Review, 56, pp. 2057–2109. doi:10.1007/s10462-022-10226-0.

Papanicolaou, A., Fu, H., Krishnamurthy, P., Healy, B. and Khorrami, F. (2023) 'An Optimal Control Strategy for Execution of Large Stock Orders Using LSTMs', arXiv preprint arXiv:2301.09705.

Rollinger, T.N. and Hoffman, S.T. (2023) Sortino: A "Sharper" Ratio. Red Rock Capital.

Sami, H. et al. (2023) 'Determining the Best Activation Functions for Predicting Stock Prices in Different Stock Exchanges', Journal of Financial Data Science, 5(2), pp. 77–94.

Schaff, D. (2008) Releasing the Code to the Schaff Trend Cycle. Technical note, 15 February.

Schulman, J. et al. (2017) 'Proximal Policy Optimization Algorithms', *arXiv preprint*, arXiv:1707.06347.

Sewak, M. (2019) *Deep Q-Network (DQN), Double DQN, and Dueling DQN.* Springer Briefs in AI.

Shavandi, A. and Khedmati, M. (2022) 'A Multi-Agent Deep Reinforcement Learning Framework for Algorithmic Trading in Financial Markets', Expert Systems with Applications, 208, 118124. doi:10.1016/j.eswa.2022.118124.

Silver, D., Singh, S., Precup, D. and Sutton, R.S. (2021) 'Reward Is Enough', Artificial Intelligence, 299, 103535.

Sutton, R.S. and Barto, A.G. (2018) *Reinforcement Learning: An Introduction (2nd ed.).* MIT Press.

Théate, T. and Ernst, D. (2021) 'An Application of Deep Reinforcement Learning to Algorithmic Trading', *Expert Systems with Applications*, 173, pp. 114–890.

Tsantekidis, A. et al. (2021) 'Price Trailing for Financial Trading Using Deep Reinforcement Learning', *IEEE Transactions on Neural Networks and Learning Systems*, 32(12), pp. 5510–5522.

van Hasselt, H., Guez, A. and Silver, D. (2016) Deep Reinforcement Learning with Double Q-Learning. arXiv preprint arXiv:1509.06461.

Van Houdt, G., Mosquera, C. and Nápoles, G. (2020) 'A review on the long short-term memory model', *Artificial Intelligence Review*, 53, pp. 5929–5955.

Wang, Z. et al. (2016) 'Dueling Network Architectures for Deep Reinforcement Learning', *arXiv preprint*, arXiv:1511.06581.

Watkins, C.J.C.H. (1989) Learning from Delayed Rewards. PhD thesis. King's College, University of Cambridge.

Xiaodong, L. et al. (2022) 'Hierarchical Deep Reinforcement Learning for VWAP Strategy Optimization', *Applied Intelligence*, 52(5), pp. 5470–5485.

Zhang, Y. et al. (2020) 'Stock Market Prediction with Multi-Agent Reinforcement Learning', *Expert Systems with Applications*, 152, 113–142.

Καϊπής, Λ. (2025) Δημιουργία αλγορίθμου για αυτόματη αγοραπωλησία μετοχών μέσα στην ημέρα για το δείκτη S&P 500. Διπλωματική εργασία. Ελληνικό Ανοικτό Πανεπιστήμιο, Τμήμα Πληροφορικής.

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.