Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Έλεγχος Συστήματος Αποθήκευσης Ενέργειας με χρήση Μεθόδων Ενισχυτικής Μάθησης

Διπλωματικη Εργασια

του

ΣΤΑΜΟΥΛΗ ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων: Εμμανουήλ Βαρβαρίγος
Καθηγητής

Αθήνα, Νοέμβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Έλεγχος Συστήματος Αποθήκευσης Ενέργειας με χρήση Μεθόδων Ενισχυτικής Μάθησης

Διπλωματικη Εργασια

του

ΣΤΑΜΟΥΛΗ ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων: Εμμανουήλ Βαρβαρίγος
Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26α Νοεμβρίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Εμμανουήλ Βαρβαρίγος	Θεοδώρα Βαρβαρίγου	Ηρακλής Αβραμόπουλος
Καθηγητής	Καθηγήτρια	Καθηγητής

Αθήνα, Νοέμβριος 2021


Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

(Υπογραφή)

.....
Κωνσταντίνος Σταμούλης

26 Σεπτεμβρίου 2021

Περίληψη

Η μεγάλη διείσδυση των Ανανεώσιμων Πηγών Ενέργειας στις αγορές τα τελευταία χρόνια απαιτεί εξίσου μεγάλη ευελιξία από τα δίκτυα ενέργειας για να μπορέσουν να εξισορροπήσουν την αστάθεια και τους κινδύνους που προκαλούν σε αυτά. Η μεγάλη ευελιξία που παρέχουν τα Συστήματα Αποθήκευσης Ενέργειας (Energy Storage Systems), σε συνδυασμό με την συνεχή μείωση του κόστους κατασκευής τους, έχουν ευνοήσει την ενσωμάτωσή τους στα δίκτυα, με σκοπό την αντιμετώπιση αυτών των ανισορροπιών.

Στόχος της παρούσας διπλωματικής εργασίας είναι η μελέτη, ο σχεδιασμός και η αξιολόγηση της απόδοσης μεθόδων Μηχανικής Μάθησης (Machine Learning) για έναν οικονομικά βέλτιστο έλεγχο ενός Συστήματος Αποθήκευσης Ενέργειας που συμμετέχει στην εξισορρόπηση της ηλεκτρικής ενέργειας σε ένα δίκτυο. Με την εφαρμογή των αλγορίθμων, λαμβάνονται υπόψη οι τιμές ανισορροπίας και μελετάται η ποσότητα ενέργειας που πρέπει να αποθηκεύσει στο σύστημα ο χειριστής, ή αντίστοιχα η ποσότητα που πρέπει να διοχετεύσει στο δίκτυο, με στόχο πάντα να μεγιστοποιήσει το τελικό κέρδος.

Πιο συγκεκριμένα το πρόβλημα μοντελοποιήθηκε ως μια Μαρκοβιανή Διαδικασία Αποφάσεων (MDP) και αναπτύχθηκαν μέθοδοι Ενισχυτικής Μάθησης για τον έλεγχο του συστήματος. Από τα αποτελέσματα προκύπτει ότι η απόδοση που προσφέρουν είναι αρκετά υψηλότερη έναντι μιας κοινής προσέγγισης που θα μπορούσε να έχει το πρόβλημα.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Ενισχυτική Μάθηση, Σύστημα Αποθήκευσης Ενέργειας, Ανισορροπία, Αγορές Ενέργειας, Αποθήκευση

Abstract

The high penetration of Renewable Energy Sources (RES) in recent years requires high flexibility in order to control the imbalances and the risks in the electrical grid. Energy storage systems (ESSs) are a promising technology due to their inherent distributed nature and their high ramping capability. While their cost is decreasing over years, it is a viable approach to optimize the exploitation of such devices.

The control of an ESS in order to maximise the storage exploitation in the presence of imbalance uncertainty results in a great challenge for system operators. The goal of this thesis is to investigate the way in which Machine Learning algorithms can be leveraged for the optimization of the flexibility offered by ESSs for reducing grid imbalances and provide the best performance to market operator from economical point of view.

In particular we focus on the development and the evaluation of an Energy Storage System that function under Reinforcement Learning methods that is able to control the ESS. The problem of the Energy Storage System control is formulated as a Markov Decision Process (MDP). The applied algorithms are compared with a baseline policy of making decisions. The results show that Reinforcement Learning algorithms outperform the baseline policy.

Keywords

Machine Learning, Reinforcement Learning, Energy Storage System, Imbalance, Energy Markets, Storage

στην οικογένειά μου

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα αρχικά να ευχαριστήσω τον κ. Βαρβαρίγο για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα. Επίσης ευχαριστώ ιδιαίτερα τον Ιωάννη Μπούκα για τις πολύτιμες συμβουλές του, την υπομονή του και την απαράμιλλη καθοδήγησή του. Ακόμα ευχαριστώ τον κ. Τσαούσογλου για όλες τις υποδείξεις και παρατηρήσεις του πάνω στην εργασία.

Στην συνέχεια, θα ήθελα να ευχαριστήσω τους φίλους μου που είναι δίπλα μου και με βοηθούν να πετυχαίνω τους στόχους μου. Κλείνοντας, θα ήθελα να ευχαριστήσω τους γονείς μου και τα αδέλφια μου, για την αμέριστη συμπαράσταση και την αδιάκοπη στήριξή τους όλα αυτά τα χρόνια.

Αθήνα, Νοέμβριος 2021

Κωνσταντίνος Σταμούλης

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η κλιματική αλλαγή που καταγράφεται τα τελευταία χρόνια δεν αποτελεί μόνο ένα τεράστιο περιβαλλοντικό πρόβλημα αλλά και μια μεγάλη πρόκληση για το αναπτυξιακό μοντέλο που θα ακολουθήσει η κάθε χώρα σε παγκόσμιο επίπεδο. Οι συνεχώς αυξανόμενες εκπομπές αερίων θερμοκηπίου έχουν οδηγήσει στην επιδείνωση της υπερθέρμανσης του πλανήτη. Πιο συγκεκριμένα εκτιμάται ότι η αύξηση της θερμοκρασίας που έχει σημειωθεί από τα τέλη του 19ου αιώνα είναι μεγαλύτερη των 1.2 $^{\circ}C$ σε παγκόσμιο επίπεδο, με τα έτη 2016 και 2020 να είναι τα θερμότερα που έχουν καταγραφεί. [1] Ακόμα εκτιμάται ότι μέχρι τον επόμενο αιώνα η αύξηση της θερμοκρασίας θα φτάσει από 1.4 $^{\circ}C$ μέχρι 5.5 $^{\circ}C$. [2] Το γεγονός αυτό μπορεί να προκαλέσει ολέθριες επιπτώσεις οι οποίες σε ορισμένες περιπτώσεις να είναι μη αναστρέψιμες, όπως για παράδειγμα η καταστροφή του οικοσυστήματος. Μια σημαντική προσπάθεια για την αντιμετώπιση της αύξησης της θερμοκρασίας, σημειώθηκε με την συμφωνία του Παρισιού, όπου τέθηκε στόχος για τον περιορισμό της αύξησης της θερμοκρασίας στον 1.5 $^{\circ}C$. Ο γενικότερος στόχος αυτής της πρωτοβουλίας ήταν ο περιορισμός της εκπομπής διοξειδίου του άνθρακα και η επίτευξη της κλιματικής ουδετερότητας έως το 2050.[3]

Στο πλαίσιο της δέσμευσης της συμφωνίας του Παρισιού, η Ευρωπαϊκή Ένωση έχει χαράξει μια μακροπρόθεσμη ενεργειακή πολιτική και έχει προχωρήσει στην αναθεώρηση του θεσμικού πλαισίου της, μέσω του προγράμματος *Clean Energy for all Europeans* [4]. Πιο συγκεκριμένα στοχεύει :

- στην βελτιστοποίηση της ενεργειακής απόδοσης των κτιρίων και την μείωση της ενεργειακής κατανάλωσης κατά 32.5% μέχρι το 2030.
- στην αύξηση του ελάχιστου μεριδίου ανανεώσιμων πηγών ενέργειας, στο 32% του ενεργειακού μίγματος της Ευρωπαϊκής Ένωσης μέχρι το 2030.
- στον συντονισμό της σχεδίασης Εθνικών ενεργειακών πολιτικών από το κάθε κράτος-μέλος με σκοπό την συμμόρφωσή τους, στους στόχους της συμφωνίας.
- στην παροχή μεγαλύτερης ευελιξίας και ασφάλειας στον καταναλωτή.
- στην θέσπιση νόμων για την λειτουργία μιας ασφαλής και αποδοτικής αγοράς ενέργειας.

Το Ευρωπαϊκό Δίκτυο Διαχειριστών Συστημάτων Μεταφοράς Ηλεκτρικής Ενέργειας (European Network of Transmission System Operators for Electricity-ENTSO-E) αφού εξετάσει την Εθνική ενεργειακή πολιτική του κάθε κράτους καθώς και αν αυτά σαν σύνολο συγκλίνουν

τους στόχους της συμφωνίας, αποφασίζει για την πορεία της μελλοντικής ενεργειακής στρατηγικής. Για την ανάπτυξη αυτής της πολιτικής, εκτιμάται τόσο η εξέλιξη της ζήτησης και της προσφοράς ενέργειας, όσο και η εκπομπή διοξειδίου του άνθρακα μέχρι το 2050. Από την παραπάνω εκτίμηση, προκύπτει ότι μέχρι το 2030 η κάλυψη των Ανανεώσιμων Πηγών Ενέργειας(ΑΠΕ) στην Ευρώπη θα έχει ξεπεράσει το 40%. Παρόμοια κατεύθυνση φαίνεται να επικρατεί και σε παγκόσμιο επίπεδο καθώς εικάζεται σύμφωνα με τον Διεθνή Οργανισμό Ανανεώσιμων Πηγών Ενέργειας (International Renewable Energy Agency-IRENA) ότι το ποσοστό της κάλυψης των ΑΠΕ θα έχει φτάσει τουλάχιστον στο 38% μέχρι το 2030.[5]

Η παραγωγή από ΑΠΕ λόγω της εξάρτησής τους από τις καιρικές συνθήκες παρουσιάζει έντονη χρονική διακύμανση με αποτέλεσμα, η πρόβλεψή της να είναι δύσκολο να επιτευχθεί. Έτσι, ένα συχνό πρόβλημα που δημιουργείται είναι η εκτιμώμενη προσφορά να μην καλύπτει την ζήτηση σε πραγματικό χρόνο, με αποτέλεσμα να μην ικανοποιείται το ενεργειακό ισοζύγιο της ισχύος. Αυτό το πρόβλημα θα γίνει ακόμα πιο έντονο όταν οι ΑΠΕ καταλάβουν μεγάλο μερίδιο της παραγωγής σε ένα δίκτυο. Η ενσωμάτωσή τους θα προσθέσει μεγάλη αστάθεια στο δίκτυο, κάνοντας την ασφάλεια και την ομαλή λειτουργία του, ένα μεγάλο πρόβλημα. Έτσι, η ευελιξία του δικτύου θα είναι ο βασικός παράγοντας για την περαιτέρω ανάπτυξη και ενσωμάτωση των ΑΠΕ σε αυτό. Ένα ευέλικτο δίκτυο μπορεί να προσαρμοστεί στις μεταβολές που προκαλούνται και να καλύψει τις αποκλίσεις στο ισοζύγιο, εξασφαλίζοντας την σταθερή και ασφαλή λειτουργία του.

Τις τελευταίες δεκαετίες, τα κόστη κατασκευής και λειτουργίας των τεχνολογιών αποθήκευσης ενέργειας και κυρίως των ιόντων λιθίου, έχουν μειωθεί σε μεγάλο βαθμό και αναμένεται να συνεχίσουν την καθοδική πτώση για τα επόμενα τριάντα χρόνια. [6][7] Αυτό, σε συνδυασμό με τον γρήγορο χρόνο απόκρισης τους κατά την λειτουργία τους, τις καθιστά ως μια ελκυστική λύση για να προσφέρουν την απαραίτητη ευελιξία στα προβλήματα που δημιουργεί η περαιτέρω ενσωμάτωση των ΑΠΕ.

Η χρήση αποθηκευτικών συστημάτων μπορεί να προσφέρει ευελιξία με πολλούς τρόπους στο δίκτυο, ανάλογα με τα χαρακτηριστικά, τον τύπο, την τεχνολογία και την χωρητικότητα που διαθέτουν.[8] Πιο συγκεκριμένα, τα συστήματα αποθήκευσης μεγάλης κλίμακας, έχουν την δυνατότητα να αποθηκεύουν ηλεκτρική ενέργεια όταν αυτή βρίσκεται σε περίσσεια και να την διοχετεύουν στο δίκτυο όταν υπάρχει υψηλή ζήτηση σε ώρα αιχμής. Μία τέτοια διαδικασία θα μπορούσε να εξισορροπήσει τα αυξημένα κόστη της ενέργειας αποφέροντας οικονομική ελάφρυνση στους τελικούς καταναλωτές.

1.1 Αντικείμενο της διπλωματικής

Εκτός των πολλών διαφορετικών και σημαντικών εφαρμογών των αποθηκευτικών συστημάτων ηλεκτρικής ενέργειας, σε αυτήν την διπλωματική μελετήθηκε η αξιοποίησή τους, με σκοπό την μελλοντική εκμετάλλευση της αποθηκευμένης ενέργειας για την προσφορά οικονομικού οφέλους είτε στον Διαχειριστή του ηλεκτρικού δικτύου είτε σε έναν φορέα που συμμετέχει ως aggregator στην αγορά εξισορρόπησης. Στόχος είναι ο διαχειριστής του Συστήματος Αποθήκευσης να εκμεταλλεύεται την μεγάλη διακύμανση των τιμών ενέργειας που προκύπτουν από τις ανισορροπίες σε μια αγορά. Οι αποφάσεις που θα παίρνει κάθε φορά ο διαχειριστής αποτελεί ένα αρκετά πολύπλοκο πρόβλημα καθώς υπάρχει μεγάλη αβεβαιότητα στην πορεία που θα ακολουθήσει η ανισορροπία σε ένα δίκτυο. Έτσι στην παρούσα διπλωματική εξετάζεται η ικανότητα διαφορετικών αλγορίθμων Ενισχυτικής Μάθησης για την βέλτιστη διαχείριση ενός συστήματος αποθήκευσης. Το σύστημα αυτό ανταποκρίνεται με βάση τις τιμές ανισορροπίας που προκύπτουν από μια αγορά ενέργειας και επιφέρει στον διαχειριστή που λαμβάνει τις αποφάσεις τις χρηματικές απολαβές από την λειτουργία του. Στόχος είναι πάντα η βέλτιστη αξιοποίησή του για να ελαχιστοποιήσει τα κόστη και να μεγιστοποιήσει το τελικό κέρδος.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό αναλύονται οι αρχές λειτουργίας των μεθόδων Ενισχυτικής Μάθησης που θα χρησιμοποιηθούν για τις ανάγκες της διπλωματικής εργασίας.

2.1 Μηχανική Μάθηση

Με τον όρο Μηχανική Μάθηση (Machine Learning-ML) εννοείται το πεδίο της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI) που ασχολείται με την μελέτη και τον σχεδιασμό υπολογιστικών συστημάτων ικανών να χρησιμοποιούν εμπειρία και γνώση για την λήψη αποφάσεων. Η Μηχανική Μάθηση εφαρμόζεται σε σύνθετα προβλήματα στα οποία οι αλγόριθμοι αδυνατούν να βρουν μία πλήρη και αποδοτική λύση. Έτσι ο σκοπός της είναι, η επεξεργασία δεδομένων με τις κατάλληλες τεχνικές και αλγορίθμους για την εκπαίδευση ενός υπολογιστικού συστήματος που θα επιλύει τέτοιου είδους προβλήματα. Οι αλγόριθμοι Μηχανικής Μάθησης αποτελούνται από τρεις διαφορετικές κατηγορίες ανάλογα με το είδος του προβλήματος προς επίλυση. Πιο συγκεκριμένα χωρίζεται σε Μάθηση με Επίβλεψη (Supervised Learning), Μάθηση χωρίς Επίβλεψη (Unsupervised Learning) και Ενισχυτική Μάθηση (Reinforcement Learning). Ορισμένες από τις πιο κοινές εφαρμογές της Μηχανικής Μάθησης είναι στην οικονομία, την ιατρική, στην αναγνώριση προτύπων και ήχου, στην επεξεργασία φυσικής γλώσσας κτλπ.[9][10][11]

2.1.1 Μάθηση με Επίβλεψη

Κατά την Μάθηση με Επίβλεψη (Supervised Learning) οι αλγόριθμοι εκπαιδεύονται με σκοπό την δημιουργία ενός μοντέλου που συνδέει τις εισόδους σε γνωστές επιθυμητές εξόδους. Προσπαθεί δηλαδή να βρει μια συνάρτηση σύνδεσης μεταξύ εισόδου-εξόδου, χρησιμοποιώντας ήδη γνωστά ζεύγη, ως κατευθυντήρια παραδείγματα. Τα παραδείγματα αυτά, για κάθε δεδομένη είσοδο έχουν αντιστοιχία με μια ετικέτα (label) που εκφράζει την επιθυμητή έξοδο για την συγκεκριμένη είσοδο. Ιδανικά ένα εκπαιδευμένο μοντέλο Μάθησης με Επίβλεψη είναι σε θέση να καθορίσει το σωστό γνωστό πρότυπο που ανήκει μια οποιαδήποτε άγνωστη νέα είσοδος. Τα προβλήματα επιβλεπόμενης μάθησης διακρίνονται σε δύο είδη:

- Ταξινόμησης (Classification), όπου δημιουργούνται μοντέλα πρόβλεψης διακριτών κατηγοριών.
- Παρεμβολής (Regression), αφορά μοντέλα πρόβλεψης συνεχών αριθμητικών τιμών.

2.1.2 Μη Επιβλεπόμενη Μάθηση

Οι αλγόριθμοι Μη Επιβλεπόμενης Μάθησης (Unsupervised Learning) έχουν στόχο την εύρεση κάποιας συσχέτισης των δεδομένων με κάποιο κανόνα ή την συγκρότηση ομάδων από αυτά. Η βασική διαφορά σε αυτήν την κατηγορία είναι ότι τα αποτελέσματα βασίζονται μόνο στις ιδιότητες των δεδομένων, χωρίς να συνοδεύονται αυτά από καμία ετικέτα.

2.1.3 Ενισχυτική Μάθηση

Η Ενισχυτική Μάθηση (Reinforcement Learning) αποτελεί ένα γενικό όρο που εκφράζει μια σειρά από τεχνικές στις οποίες το σύστημα μάθησης εκπαιδεύεται μέσω της συνεχόμενης αλληλεπίδρασης με το περιβάλλον του. Αποτελεί ένα κλάδο της Μηχανικής Μάθησης που ασχολείται με την εύρεση ενεργειών για τον βέλτιστο έλεγχο ενός προβλήματος. Σε αυτήν την περίπτωση, κατά την διαδικασία της εκπαίδευσης δεν παρέχεται εξ΄ αρχής η πληροφορία και η γνώση για το ποιες ενέργειες είναι προτιμότερες, αλλά το σύστημα τις ανακαλύπτει μόνο του μέσω της αλληλεπίδρασης. Σκοπός του συστήματος μάθησης είναι να μεγιστοποιήσει μια συνάρτηση του αριθμητικού σήματος ενίσχυσης (ανταμοιβή), με στόχο να αναπτύξει μια βέλτιστη στρατηγική. Αποτελεί ίσως το πιο κοινό είδος μάθησης στον πραγματικό κόσμο, αφού είναι εμπνευσμένη από τα αντίστοιχα ανάλογα της δοκιμής και της αποτυχίας (trial and error), που συναντώνται ως τρόποι μάθησης στα έμβια όντα. [12]

2.2 Μοντελοποίηση Προβλήματος Λήψης Αποφάσεων

Η Ενισχυτική Μάθηση εφαρμόζεται στην μελέτη και την βελτιστοποίηση προβλημάτων λήψης αποφάσεων. Τα προβλήματα αυτά μοντελοποιούνται συνήθως ως μία Μαρκοβιανή Διαδικασία Απόφασης (Markov Decision Process), η οποία αποτελεί μία στοχαστική διαδικασία ελέγχου διακριτού χρόνου. Μια MDP στηρίζεται σε μια ακολουθία μετάβασεων κατάστασης μετά από την επιλογή ορισμένων ενεργειών. Ο στόχος της είναι η ανάπτυξη μιας πολιτικής που λύνει την MDP και ταυτόχρονα μεγιστοποιεί την μακροπρόθεσμη ανταμοιβή. Στην συνέχεια του κεφαλαίου, αναλύεται με περισσότερες λεπτομέρειες.

2.2.1 Μαρκοβιανή Διαδικασία Αποφάσεων (Markov Decision Process)

Μια MDP είναι μια μαθηματική έκφραση για την αναπαράσταση των στοχαστικών διαδικασιών λήψης διαδοχικών αποφάσεων για διακριτές τιμές του χρόνου[13]. Σε κάθε χρονική στιγμή $t = 0, 1, 2, 3, ...$ η διαδικασία βρίσκεται σε κάποια κατάσταση S_t, στην οποία μπορεί να επιλεχθεί οποιαδήποτε ενέργεια A_t είναι δυνατή στην δεδομένη κατάσταση. Μετά την εκτέλεση της ενέργειας η διαδικασία θα βρίσκεται σε μια νέα κατάσταση S_{t+1} με μια πιθανότητα η οποία δίνεται απο την συνάρτηση μεταβάσεων. Η συνάρτηση αυτή εξαρτάται και παίρνει τιμές από τον συνδυασμό της ενέργειας A_t στην κατάσταση S_t. Ως αποτέλεσμα της διαδικασίας επιστρέφεται η ανάλογη ανταμοιβή. Πιο συγκεκριμένα μια MDP ορίζεται ως μια πλειάδα τεσσάρων στοιχείων (S, A, P, R) όπου:

- S : το σύνολο με όλες τις δυνατές καταστάσεις. Η S_t, είναι η κατάσταση που βρίσκεται το σύστημα μάθησης την χρονική στιγμή t.
  Ισχύει ότι $S_t \in S$.

- A : το πλήθος των δυνατών ενεργειών. Η A_t, είναι η ενέργεια που επιλέγει το σύστημα μάθησης την χρονική στιγμή t.
  $A_t \in A$.

- P : η συνάρτηση μεταβάσεων για κάθε ζεύγος (S_t, A_t). Για μια δεδομένη κατάσταση S_t και ενέργεια A_t την χρονική στιγμή t, η πιθανότητα να γίνει μετάβαση στην επόμενη κατάσταση S_{t+1} είναι :

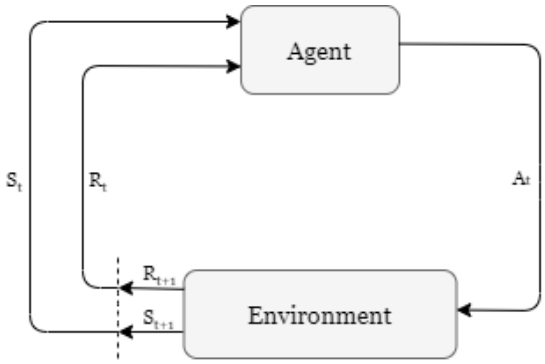
$$P(S_t)^a_{SS'} = P(S_{t+1} = S' | S_t = S, A_t = a)$$

- R : η ανταμοιβή που αντιστοιχεί στον συνδυασμό της ενεργειας A_t και την μετάβαση στην νέα κατάσταση S_{t+1}.
  $R_a \in R$, R: Σύνολο ανταμοιβών

Η βασική ιδιότητα στις Μαρκοβιανές διαδικασίες είναι ότι δεν διατηρούν μνήμη για τις προηγούμενες μεταβολές των καταστάσεων τους. Έτσι κάθε επόμενη κατάσταση εξαρτάται αποκλειστικά και μόνο από την τωρινή χωρίς να επηρεάζεται από τις προηγούμενες.

2.2.2 Πράκτορας (Agent) - Περιβάλλον (Environment)

Για να μετατραπεί το πρόβλημα προς εξέταση σε μια ΜΔΠ πρέπει να καθοριστούν οι έννοιες του Περιβάλλοντος (Environment) και του Πράκτορα (Agent). Πιο συγκεκριμένα, καλούμε πράκτορα αυτόν που μαθαίνει και παίρνει τις αποφάσεις για την κάθε ενέργεια και περιβάλλον οτιδήποτε άλλο με το οποίο αλληλεπιδρά ο Πράκτορας για την εκμάθησή του. Από την συνεχόμενη αλληλεπίδρασή τους, ο Πράκτορας επιλέγει κάθε φορά μια ενέργεια και το Περιβάλλον προσαρμόζει την νέα κατάσταση που θα προκύψει αποδίδοντας παράλληλα τις ανάλογες ανταμοιβές της κάθε ενέργειας. Στόχος του Πράκτορα είναι να μεγιστοποιήσει τις ανταμοιβές που λαμβάνει σε βάθος χρόνου με βάση τις ενέργειες που λαμβάνει.



Σχήμα 2.1: *Ενισχυτική Μάθηση*

Πιο συγκεκριμένα ο Πράκτορας και το Περιβάλλον αλληλεπιδρούν σε μια αλληλουχία διακριτών χρονικών στιγμών (discrete time steps) $t = 0, 1, 2, 3, ...$ Σε κάθε χρονική στιγμή, ο Πράκτορας λαμβάνει απεικονίσεις των καταστάσεων του περιβάλλοντος, $S_t \in S$ και επιλέγει μια ενέργεια $A_t \in A(s)$. Την επόμενη χρονική στιγμή ως αποτέλεσμα της ενέργειας αυτής λαμβάνει μια αριθμητική ανταμοιβή $R_{t+1} \in R$ και καταλήγει στην νέα κατάσταση S_{t+1}. Με αυτόν τον τρόπο η ΜΔΠ και ο Πράκτορας δημιουργούν μια ακολουθία ως εξής:

$$S_0, A_0, R_0, S_1, A_1, R_1, S_2, A_2, R_2, S_3, A_3, R_3...$$

Σε μια πεπερασμένη MDP, τα σύνολα των καταστάσεων, ενεργειών και ανταμοιβών (S, A και R) έχουν όλα πεπερασμένο αριθμό στοιχείων. Σε αυτήν την περίπτωση, οι τυχαίες μεταβλητές R_t και S_t παίρνουν τιμές από διακριτές κατανομές πιθανότητας οι οποίες εξαρτώνται μόνο από την προηγούμενη κατάσταση και ενέργεια. Δηλαδή για συγκεκριμένες τιμές αυτών των τυχαίων μεταβλητών, $s' \in S$ και $r \in R$, η πιθανότητα να συμβούν την χρονική στιγμή t, δεδομένου της προηγούμενης κατάστασης και ενέργειας ορίζεται ως:

$$p(s', r|s, a) \doteq Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a$$

για όλα τα $s', s \in S, r \in R$ και $a \in A(s)$. Η συνάρτηση $p : S \times R \times S \times A \to [0, 1]$ είναι μια συνήθης ντετερμινιστική συνάρτηση τεσσάρων μεταβλητών όπου ορίζεται από την κατανομή πιθανότητας για κάθε επιλογή των s και a, δηλαδή:

$$\sum_{s' \in S} \sum_{r' \in R} p(s', r|s, a) = 1$$

για όλα τα $s \in S$ και $a \in A(s)$. Οι πιθανότητες της συνάρτησης p των τεσσάρων στοιχείων εκφράζουν πλήρως μια πεπερασμένη MDP. Με βάση τα παραπάνω μπορούν εύκολα να οριστούν οι συναρτήσεις για την λειτουργία του περιβάλλοντος. Πιο συγκεκριμένα για τις εναλλαγές των καταστάσεων ορίζονται οι συναρτήσεις μετάβασης των καταστάσεων (state-transition probabilities) με μια μικρή παραλλαγή της συνάρτησης p, με $p : S \times S \times A \to [0, 1]$ όπου σε αυτήν την περίπτωση θα ορίζεται ως

$$p(s'|s, a) \doteq Pr\{S_t = s'|S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in R} p(s', r|s, a)$$

Αντίστοιχα οι αναμενόμενες ανταμοιβές για τα ζεύγη κατάσταση-ενέργεια(S, A) η r, όπου $r : S \times A \to \mathbb{R}$ ορίζεται:

$$r(s, a) \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R} r \sum_{s' \in S} p(s', r|s, a)$$

ενώ για τα ζεύγη κατάσταση-ενέργεια-κατάσταση(S, A, S) η r όπου $r : S \times A \times S \to \mathbb{R}$ ορίζεται:

$$r(s, a, s') \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in R} r \frac{p(s', r|s, a)}{p(s'|s, a)}$$

Σύμφωνα με τον γενικό κανόνα οτιδήποτε δεν μπορεί να ελεγχθεί από τον Πράκτορα βρίσκεται

εκτός αυτού και ανήκει στο Περιβάλλον. Δεν είναι υποχρεωτικό, το περιβάλλον να είναι τελείως άγνωστο για τον Πράκτορα, αντιθέτως σε ορισμένες περιπτώσεις μπορεί να έχει πλήρη γνώση για την λειτουργία του περιβάλλοντος αλλά να αντιμετωπίζει μεγάλες δυσκολίες στην επίλυση του προβλήματος.

2.2.3 Ανταμοιβές και Επεισόδια

Στην Ενισχυτική Μάθηση ο στόχος του Πράκτορα είναι να μεγιστοποιεί την αθροιστική ανταμοιβή (Reward) που λαμβάνει από το περιβάλλον ύστερα από μια αλληλουχία αποφάσεων που επέλεξε. Δεδομένου ότι η ανταμοιβή του για μια χρονική στιγμή t ισοδυναμεί με R_t τότε ο στόχος του μπορεί να εκφραστεί ως η μεγιστοποίηση της αναμενόμενης τιμής του συνολικού αθροίσματος των R_t, δηλαδή της ποσότητας: $G_t = R_{t+1} + R_{t+2} + ... + R_T$ όπου την χρονική στιγμή $t = T$, η αλληλουχία έχει φτάσει σε μια έννοια τερματισμού, καταλήγει δηλαδή ο Πράκτορας στην τελική κατάσταση (terminal state) του περιβάλλοντος. Κάθε φορά που ολοκληρώνεται μια τέτοια διαδικασία, ολοκληρώνεται ένα επεισόδιο (episode).

2.2.4 Πεπερασμένος - Άπειρος Ορίζοντας

Τα προβλήματα τα οποία έχουν διάρκεια για ένα σταθερό χρονικό διάστημα H και οι ενέργειες λαμβάνονται για τις χρονικές στιγμές $t = 1, ..., H$ ονομάζονται πεπερασμένου ορίζοντα (finite-horizon). Σε αυτήν την περίπτωση το κριτήριο για την συνολική ανταμοιβή είναι η μεγιστοποίηση της αναμενόμενης τιμής του αθροίσματος όλων των ανταμοιβών συνολικής διάρκειας H. Ο στόχος είναι να μεγιστοποιηθεί η αναμενόμενη τιμή του αθροίσματος των ανταμοιβών, δηλαδή η ποσότητα:

$$\mathbb{E}[\sum_{t=1}^{H} r_t]$$

Όταν η εκμάθηση και η λήψη αποφάσεων δεν τερματίζει μετά από ένα πεπερασμένο χρονικό διάστημα, τα προβλήματα ονομάζονται απείρου ορίζοντα (infinte-horizon). Σε αυτήν την περίπτωση εφαρμόζεται ο παράγοντας μείωσης γ, ο οποίος παίρνει τιμές $0 \leq \gamma \leq 1$ και ο στόχος είναι να μεγιστοποιηθεί η αναμενόμενη τιμή του αθροίσματος των σταθμευμένων από τον παράγοντα άμεσων ανταμοιβών, δηλαδή η ποσότητα:

$$\mathbb{E}[\sum_{t=1}^{H} \gamma^t r_t]$$

2.2.5 Πολιτική (Policy)

Με τον όρο πολιτική (policy) εννοούμε την αντιστοιχία των καταστάσεων με τις πιθανότητες να επιλεγεί κάθε πιθανή ενέργεια. Σε περίπτωση που ο Πράκτορας ακολουθεί μια πολιτική π σε μια χρονική στιγμή t, τότε η $\pi(\alpha|s)$ είναι η πιθανότητα για $A_t = a$ δεδομένου ότι $S_t = s$. Όπως η p έτσι και η π είναι μια κοινή συνάρτηση που εκφράζει κατανομές πιθανότητας. Οι μέθοδοι της Ενισχυτικής Μάθησης καθορίζουν την μεταβολή των πολιτικών του Πράκτορα ως αποτέλεσμα της εμπειρίας του.

Η βέλτιστη πολιτική για προβλήματα πεπερασμένου ορίζοντα ονομάζεται μη στάσιμη (non stationary) καθώς κάθε βέλτιστη ενέργεια σε μια δεδομένη κατάσταση μπορεί να αλλάζει ανάλογα με τον χρόνο. Αν όμως δεν υπάρχει σταθερό όριο χρόνου, δεν υπάρχει λόγος να έχουμε διαφορετική συμπεριφορά στην ίδια κατάσταση σε διαφορετικούς χρόνους. Έτσι η βέλτιστη ενέργεια εξαρτάται μόνο από την τρέχουσα κατάσταση, και η βέλτιστη πολιτική είναι στάσιμη (stationary).Έτσι οι πολιτικές για την περίπτωση του άπειρου ορίζοντα είναι απλούστερες από εκείνες για την περίπτωση του πεπερασμένου ορίζοντα.

2.2.6 Value Functions

Σχεδόν όλοι οι αλγόριθμοι Ενισχυτικής Μάθησης περιλαμβάνουν συναρτήσεις τιμών (*value functions*), δηλαδή συναρτήσεις καταστάσεων (ή συναρτήσεις ζεύγους καταστάσεων-ενεργειών) που υπολογίζουν την απόδοση για τον Πράκτορα να βρίσκεται κάθε φορά στη συγκεκριμένη κατάσταση (ή την απόδοση να επιλέξει μια ενέργεια σε κάποια κατάσταση). Ο όρος ¨απόδοση' εκφράζει τις μελλοντικές ανταμοιβές που μπορεί να αναμένει ο Πράκτορας. Όπως είναι λογικό αυτές οι ανταμοιβές που θα λάβει μελλοντικά εξαρτώνται από τις ενέργειες που θα επιλέξει. Αναλόγως οι *value functions* προκύπτουν από συγκεκριμένες επιλογές ενεργειών,δηλαδή ακολουθώντας συγκεκριμένη πολιτική (policy).

Η τιμή (value) μιας κατάστασης s υπό μια πολιτική π, συμβολίζεται ως $u_\pi(s)$ και ισούται με την αναμενόμενη τιμή του αποτελέσματος με αφετηρία την κατάσταση s ακολουθώντας την πολιτική π. Για μια MDP η συνάρτηση της $u_\pi(s)$ ορίζεται ως:

$$u_\pi(s) \doteq \mathbb{E}_\pi\big[\, G_t | S_t = s\big] = \mathbb{E}_\pi\big[\, \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\big]$$

για όλες τις $s \in S$, όπου $\mathbb{E}_\pi$ δηλώνει την αναμενόμενη τιμή μιας τυχαίας μεταβλητής δεδομένου ότι ο Αγεντ ακολουθεί την πολιτική π και t η χρονική στιγμή. Για κάθε τελική κατάσταση η $value function$ μηδενίζεται.

Παρομοίως ορίζεται και η συνάρτηση $q_\pi(s, a)$ για την εκτίμηση της εκτέλεσης της ενέργειας a από την κατάσταση s υπό την πολιτική π (action-value). Η συνάρτηση αυτή ορίζεται ως

$$q_\pi(s, a) \doteq \mathbb{E}_\pi\big[\, G_t | S_t = s, A_t = a\big] = \mathbb{E}_\pi\big[\, \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\big]$$

Μια βασική ιδιότητα των συναρτήσεων $u_\pi(s), q_\pi(s, a)$ είναι ότι ικανοποιούν τις αναδρομικές σχέσεις

2.3 Μέθοδοι Ενισχυτικής Μάθησης

2.3.1 Δυναμικός Προγραμματισμός

Ο Δυναμικός Προγραμματισμός (Dynamic Programming-DP) αναφέρεται στο σύνολο των αλγορίθμων οι οποίοι μπορούν να χρησιμοποιηθούν για να υπολογίσουμε βέλτιστες πολιτικές σε ένα ιδανικό μοντέλο περιβάλλοντος ως μια Μαρκοβιανή Διαδικασία Αποφάσεων(MDP). Οι κλασσικοί DP αλγόριθμοι είναι πεπερασμένης χρησιμότητας στην Ενισχυτική Μάθηση

λόγω της ανάγκης ενός ιδανικού μοντέλου περιβάλλοντος αλλά και εξαιτίας του απαγορευμένου υπολογιστικού κόστους που απαιτούν. Παρ΄ όλα αυτά εξακολουθούν να είναι σημαντικοί σε θεωρητικό επίπεδο καθώς αποτελούν την βάση για την προσέγγιση των μεθόδων που χρησιμοποιούνται στα προβλήματα της Ενισχυτικής Μάθησης. Υποθέτουμε ότι το περιβάλλον είναι μια πεπερασμένη MDP, δηλαδή τα σύνολα των καταστάσεων S, των ενεργειών A και των ανταμοιβών R είναι πεπερασμένα και οι τιμές τους δίνονται από ένα σύνολο πιθανοτήτων $p(s', r|s, a)$ για όλα τα $s \in S$ $a \in A(s)$ $r \in R$ και $s' \in S^+$ όπου S^+ το S με την τελική του κατάσταση. Παρά το γεγονός ότι η ιδέα του Δυναμικού Προγραμματισμού μπορεί να εφαρμοστεί σε προβλήματα με συνεχή χώρο καταστάσεων και ενεργειών, ακριβείς λύσεις είναι εφικτές μόνο σε ειδικές περιπτώσεις. Ο πιο συνηθισμένος τρόπος για την εύρεση προσεγγιστικών λύσεων για προβλήματα με συνεχή χώρο καταστάσεων και ενεργειών είναι η διακριτοποίηση των χώρων αυτών και στην συνέχεια η εφαρμογή των DP μεθόδων πεπερασμένων καταστάσεων. Η βασική ιδέα του DP και της ενισχυτικής μάθησης γενικότερα είναι η χρήση συναρτήσεων τιμών για την οργάνωση και την δόμηση της αναζήτησης αποδοτικών πολιτικών. Μπορούμε εύκολα να υπολογίσουμε τις βέλτιστες πολιτικές, γνωρίζοντας τις βέλτιστες συναρτήσεις τιμών u_* ή q_* οι οποίες ικανοποιούν τις ισότητες του Bellman:

$$\begin{aligned} u_*(s) &= max_\alpha \mathbb{E}\big[\, R_{t+1} + \gamma u_*(S_{t+1}) | S_t = s, A_t = a)\big] \\ &= max_\alpha \sum_{s',r} p(s', r|s, a)\big[r + \gamma u_*(s')\big] \\ q_*(s) &= \mathbb{E}\big[\, R_{t+1} + \gamma max_{\alpha'} q_*(S_{t+1}, a') | S_t = s, A_t = a)\big] \\ &= \sum_{s',r} p(s', r|s, a)\big[r + \gamma max_{\alpha'} q_*(s', a')\big] \end{aligned}$$

για όλα τα $s \in S$ $a \in A(s)$ $r \in R$ και $s' \in S^+$.

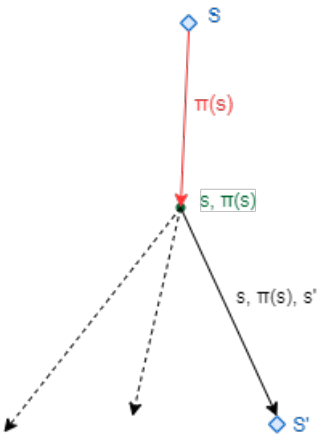
Με συνεχόμενη ανανέωση των εξισώσεων Bellman για την βελτίωση κάθε φορά της προσεγγισης της επιθυμητής συνάρτησης εκτίμησης προκύπτουν οι DP αλγόριθμοι.

2.3.1.1 Policy iteration

Δεδομένου μιας τυχαίας αρχικής πολιτικής π ο αλγόριθμος policy iteration με επαναληπτικές διαδικασίες συγκλίνει σε μια βέλτιστη π. Πιο συγκεκριμένα σε κάθε επανάληψη γίνεται αποτίμηση της πολιτικής (*policy evaluation*) με τον υπολογισμό της συνάρτησης κατάστασης-τιμών $V(s) = \sum_{s',r'} p(s', r|s, \pi(s)[r + \gamma V(s')])$, ενώ ύστερα πραγματοποιείται η βελτίωση της πολιτικής (*policy improvement*) μέσω της ανανέωσης: $\pi(s) = argmax_\alpha \sum_{s',r'} p(s', r|s, a)[r + \gamma V(s')]$. Συμβολίζοντας με $\xrightarrow{\text{E}}$ την εκτίμηση της πολιτικής (*policy evaluation*) και με $\xrightarrow{\text{I}}$ την βελτίωση (*policy improvement*) προκύπτει μια σειρά από συνεχόμενα βελτιωμένες σε σχέση με τις προηγούμενες πολιτικές και συναρτήσεις εκτίμησης όπως φαίνεται παρακάτω:

$\pi_0 \xrightarrow{\text{E}} u_{\pi 0} \xrightarrow{\text{I}} \pi \xrightarrow{\text{E}} u_{\pi 1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} u_*$

Κάθε νέα πολιτική αποτελεί βελτίωση της προηγούμενης μέχρι να γίνει η βέλτιστη. Επειδή μια πεπερασμένη MDP έχει πεπερασμένο αριθμό πολιτικών, αυτή η διαδικασία πρέπει να συγκλίνει σε μια βέλτιστη πολιτική και βέλτιστη συνάρτηση εκτίμησης σε πεπερασμένο αριθμό



Σχήμα 2.2: *Policy iteration steps diagram*

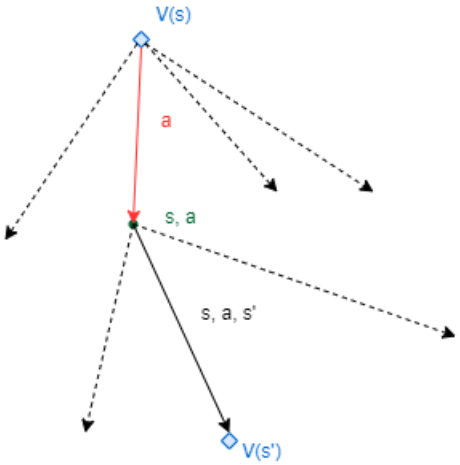
επαναλήψεων.

2.3.1.2 Value iteration

Ένα αρνητικό του policy iteration είναι ότι κάθε μια από τις επαναλήψεις του στηρίζεται στο *policy evaluation*, γεγονός που μπορεί να οδηγήσει σε εκτενείς επαναλήψεις. Μια πιο απλή περίπτωση προκύπτει όταν μελετηθεί μόνο μία επανάληψη κατά τον *policy evaluation*. Αυτή η περίπτωση ονομάζεται *value iteration* και εκφράζεται απλά με μια ανανέωση που συνδυάζει βελτίωση της πολιτικής με

$$u_{k+1}(s) \doteq max_\alpha \mathbb{E}\big[R_{t+1} + \gamma u_k(S_{t+1}) | S_t = s, A_t = a\big]$$
$$= max_\alpha \sum_{s',r} p(s', r|s, a)[r + \gamma u_k(s')]$$

για όλες τις $s \in S$. Για τυχαία u_0, η αλληλουχία u_k μπορεί να συγκλίνει στην u_* υπό τις ίδιες συνθήκες που εξασφαλίζουν την ύπαρξη της u_*.



Σχήμα 2.3: *Value iteration steps diagram*

Αλγοριθμος 2.1: *Policy Iteration*

1: Initialization
$V(s) \in R$ and $\pi(s) \in A(s)$ for all $s \in S$

2: Policy Evaluation
Repeat
$\Delta \leftarrow 0$
For each $s \in S$:
$u \leftarrow V(s)$
$V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))\big[r + \gamma V(s')\big]$
$\Delta \leftarrow max(\Delta, |u - V(s)|)$
until $\Delta < \theta$ (θ : a small positive number)

3: Policy Improvement
$police - stable \leftarrow true$
For each $s \in S$:
$old - action \leftarrow \pi(s)$
$\pi(s) \leftarrow argmax_a \sum_{s',r} p(s', r|s, \alpha)\big[r + \gamma V(s')\big]$
If $old - action \neq \pi(s)$:
$policy - stable \leftarrow false$
If $policy - stable$:
stop and return $V \approx u_*$ and $\pi \approx \pi_*$
else:
goto 2

Αλγοριθμος 2.2: *Value Iteration*

Initialize array V

Repeat
$\Delta \leftarrow 0$
For each $s \in S$:
$u \leftarrow V(s)$
$V(s) \leftarrow max_\alpha \sum_{s',r} p(s', r|s, \alpha)\big[r + \gamma V(s')\big]$
$\Delta \leftarrow max(\Delta, |u - V(s)|)$
until $\Delta < \theta$ (θ : a small number)

Return: a deterministic policy, $\pi \approx \pi_*$,
$\pi(s) = argmax_\alpha \sum_{s',r} p(s', r|s, \alpha)\big[r + \gamma V(s')\big]$.

2.3.2 Model-Free και Model-Based

Οι τεχνικές ενισχυτικής μάθησης αποτελούνται από το σύνολο των λύσεων που λαμβάνουν υπόψη τις ενέργειες που θα προσφέρουν βέλτιστα μακροχρόνια ή βραχυχρόνια αποτελέσματα. Οι προσεγγίσεις με μοντέλο (model-based) περιγράφουν με απόλυτη σαφήνεια την δομή για το περιβάλλον και τον Agent. Το μοντέλο αυτό εκφράζει τα αποτελέσματα των ενεργειών και των αντίστοιχων ανταμοιβών που προκύπτουν. Πιο συγκεκριμένα κάθε απόφαση και αντίστοιχα κάθε ανταμοιβή προκύπτει με βάση την συνάρτηση πυκνότητας πιθανότητας των μεταβάσεων και ανταμοιβών. Σε αντίθετη περίπτωση οι προσεγγίσεις χωρίς μοντέλο (model-free) δεν στηρίζονται σε κάποια συγκροτημένη δομή ενός μοντέλου αλλά κάνουν εκτίμηση των ενεργειών μέσω της δοκιμής και του σφάλματος. Ενώ οι model-free μέθοδοι δεν στηρίζονται από την δυνητική απόδοση της χρήση ενός μοντέλου, τείνουν να είναι πιο εύκολο να εφαρμοστούν. Οι μέθοδοι χωρίς μοντέλο είναι πιο δημοφιλείς και έχουν αναπτυχθεί και δοκιμαστεί εκτενέστερα από τις μεθόδους που βασίζονται σε μοντέλα.

2.3.3 On-Policy και Off-Policy

Οι επιλογές που λαμβάνει ο Agent εξαρτώνται από την πολιτική που χρησιμοποιείται σε κάθε περίπτωση. Όταν η πολιτική που ακολουθείται είναι ανεξάρτητη των επιλογών του Agent τότε πρόκειται για off-policy μέθοδο. Σε αυτήν την περίπτωση, η εύρεση της βέλτιστης πολιτικής π είναι ανεξάρτητη της διαδικασίας που ακολουθεί η εκπαίδευση. Στην περίπτωση που χρησιμοποιείται on-policy προσέγγιση γίνεται εκτίμηση και βελτίωση της ίδιας πολιτικής π με βάση την οποία ενεργεί ο Agent.

2.3.4 Αλγόριθμοι Ενισχυτικής Μάθησης

Μία από τις κατηγορίες των μεθόδων Ενισχυτικής Μάθησης είναι αυτές που στηρίζονται στον υπολογισμό των τιμών κατάστασης (state value) ή κατάστασης-ενέργειας (state-action value). Σε αυτήν την περίπτωση δεν υπάρχει εξάρτηση από κάποια ρητά καθορισμένη πολιτική, αντιθέτως αυτή προκύπτει κάθε φορά από την απευθείας επιλογή της μέγιστης τιμής κατάστασης ή που έχει υπολογιστεί. Τέτοιου είδους αλγόριθμοι προκειμένου να λειτουργήσουν με επιτυχία είναι αναγκαίο σε αρχικό στάδιο να δοκιμαστούν σε εξερευνηση του περιβάλλοντος για να ανανεωθούν κατάλληλα οι τιμές τους. Στην συνέχεια αναλύονται με περισσότερες λεπτομέρειες όλοι οι αλγόριθμοι αυτής της μορφής που χρησιμοποιήθηκαν.

2.3.4.1 Q-learning

Μία από τις βασικότερες τεχνικές της Ενισχυτικής Μάθησης, αποτελεί η τεχνική του ελεγκτικού αλγορίθμου Q-Learning, όταν δεν είναι εφικτή η εφαρμογή αλγορίθμων Δυναμικού Προγραμματισμού. Πιο συγκεκριμένα, είναι χρήσιμο να οριστεί πρώτα μία νέα συνάρτηση $Q(S_t, A_t)$, στην οποία ο Πράκτορας ξεκινάει από την κατάσταση Σ, επιλέγει την δράση α και στην συνέχεια θεωρείται ότι ενεργεί βέλτιστα. Έτσι, η συνάρτηση $Q(S_t, A_t)$ εκτιμά την τιμή ενός συνδυασμού κατάστασης και δράσης (S, A) ως εξής:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max_{\alpha} Q(S_{t+1}, \alpha) - Q(S_t, A_t)]$$

Αλγοριθμος 2.3: *Q-learning*

Initialize $Q(s,a)$, for all $s \in S, \alpha \in A(s)$ and $Q(\text{terminal-state}\cdot) = 0$
for (each episode) **do** :
    Initialize S
    repeat(for each step of episode):
        Choose A from S using a policy
        Take action A, observe R,S'
        $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma max_a Q(S',A') - Q(S,A)\big]$
        $S \leftarrow S'$
    until S is terminal
end for

- α: ο ρυθμός μάθησης, δηλαδή ο παράγοντας που καθορίζει πόσο γρήγορα ο Πράκτορας ανανεώνει την τιμή της $Q(S_t, A_t)$ με νέες τιμές και υπερισχύει έναντι των παλιών. Για την παράμετρο α ισχύει $0 < \alpha \leq 1$. Όταν η τιμή είναι 0 σημαίνει πως οι τιμές $Q(S_t, A_t)$ παραμένουν σταθερές και δεν ενημερώνονται καθ' όλη την εκτέλεση του αλγορίθμου. Αντίθετα με την τιμή 1 ο Πράκτορας αναγκάζεται να λαμβάνει υπόψη μόνο τις πιο πρόσφατες παρατηρήσεις αγνοώντας τις παρελθοντικές.

- γ: Ο παράγοντας μείωσης. Αν ο στόχος του αλγορίθμου είναι η συνεχής αναζήτηση της μέγιστης ανταμοιβής για κάθε επανάληψη είναι πιθανό να υψηλή η βραχυχρόνια επιβράβευσή του, αλλά η μακροχρόνια να μην είναι η μέγιστη δυνατή. Έτσι δεδομένου ότι η ενέργεια που αποφέρει την μέγιστη ανταμοιβή σε μια τρέχουσα κατάσταση δεν είναι αντίστοιχα και η βέλτιστη, η προσθήκη του παράγοντα μείωσης δίνει την δυνατότητα στον Πράκτορα να καθορίσει τον βαθμό εκμετάλλευσης των μελλοντικών ανταμοιβών. Για τον παράγοντα γ ισχύει $0 \leq \gamma \leq 1$

Στην παραπάνω εξίσωση, λοιπόν, ο Πράκτορας όταν βρίσκεται στην κατάσταση S_t, επιλέγει την δράση a, αυτή τον οδηγεί στην κατάσταση S_{t+1} και εκεί επιλέγει την βέλτιστη δράση α' που θα μεγιστοποιήσει την συνάρτηση $Q(S_t, A_t)$.

2.3.4.2 Double Q learning

Για την λύση στοχαστικών ΜΔΠ, ο αλγόριθμος $Q-learning$ μπορεί να αποφευχθεί λιγότερο αποδοτικός καθώς υπάρχει το ενδεχόμενο να υπερεκτιμήσει τις συναρτήσεις τιμών κατά τον υπολογισμό τους. Η συνεχόμενη χρήση της μέγιστης αναμενόμενης τιμής για τις ανανεώσεις των τιμών μπορεί να οδηγήσει σε πόλωση των τελικών αποτελεσμάτων. Με μια παραλλαγή, ο αλγόριθμος $Double\ Q-learning$, [14] πραγματοποιεί διπλή εκτίμηση χρησιμοποιώντας δύο διαφορετικές συναρτήσεις τιμών Q_1, Q_2 για ένα περισσότερο αμερόληπτο αποτέλεσμα. Κάθε Q συνάρτηση ενημερώνεται με μια τιμή από την άλλη Q συνάρτηση για την επόμενη κατάσταση. Πιο συγκεκριμένα επιλέγεται η ενέργεια α η οποία μεγιστοποιεί την συνάρτηση τιμών Q_1 για την κατάσταση S'. Όμως αντί να χρησιμοποιηθεί η τιμή που δίνει η Q_1 για την α όπως θα έκανε η μέθοδος $Q-learning$ υπολογίζεται η τιμή της Q_2 δεδομένου της α στην S' για την ανανέωση. Ανάλογη ανανέωση με τυχαία επιλογή μεταξύ των δύο

γίνεται και για την Q_2. Έτσι με την τεχνική $Double\ Q-learning$ επιτυγχάνεται ταχύτερη σύγκλιση στο βέλτιστο αποτέλεσμα.

Αλγοριθμος 2.4: *Double Q-learning*

Initialize $Q_1(s,a)$ and $Q_2(s,a)$, for all $s \in S, \alpha \in A(s)$.
Initialize $Q_1(terminal\text{-}state\cdot) = Q_2(terminal\text{-}state\cdot) = 0$
for (each episode) **do**:
    Initialize S
    repeat(for each step of episode):
        Choose A from S using a policy
        Take action A, observe R,S'
        With 0.5 probability:
            $Q_1(S,A) \leftarrow Q_1(S,A) + \alpha\big[R + \gamma Q_2(S', argmax_\alpha(Q_1(S',\alpha)) - Q_1(S,A)\big]$
        else:
            $Q_2(S,A) \leftarrow Q_2(S,A) + \alpha\big[R + \gamma Q_1(S', argmax_\alpha(Q_2(S',\alpha)) - Q_2(S,A)\big]$
        $S \leftarrow S'$
    until S is terminal
end for

2.3.4.3 SARSA

Με βάση την ίδια λογική αλλά αυτήν την φορά επιλέγοντας κάθε ενέργεια για την ανανέωση των τιμών της συνάρτησης $Q(S_t, A_t)$ δημιουργείται ένας νέος αλγόριθμος βασισμένος σε συγκεκριμένη πολιτική(on-policy). Σε πρώτο στάδιο επιλέγεται μια ενέργεια (A_{t+1}) με βάση μια πολιτικής π και προκύπτει ένας νέος συνδυασμός ζευγους S_{t+1}, A_{t+1} όπου S_{t+1} η νέα κατάσταση. Έτσι γίνεται εκτίμηση για έναν νέο συνδυασμό, πάντα με βάση την πολιτική π και η ανανέωση της $Q(S_t, A_t)$ ισοδυναμεί με :

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$$

Η παραπάνω ανανέωση πραγματοποιείται μετά από κάθε μετάβαση από μια μη τελική κατάσταση S_t. Αν η S_{t+1} είναι τελική τότε η $Q(S_{t+1}, A_{t+1})$ μηδενίζεται. Αυτός ο κανόνας χρησιμοποιεί κάθε φορά την πεντάδα $(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$ για την ανανέωση των τιμών κάθε επόμενου ζεύγους κατάστασης-ενέργειας (S_t, A_t). Από την παραπάνω πεντάδα προκύπτει το αντίστοιχο όνομα του αλγορίθμου. Η σύγκλιση του αλγορίθμου SARSA εξαρτάται από την εξάρτηση της εκάστοτε πολιτικής στην τιμή της συνάρτησης $Q(S_t, A_t)$.

2.3.4.4 Expected SARSA

Ο αλγόριθμος Expected SARSA είναι παρόμοιος με τον αλγόριθμο Q-learning. Η κύρια διαφορά έγκειται στο κριτήριο για την επιλογή της ενέργειας a, καθώς δεν είναι η μεγιστοποίηση του ζεύγους κατάστασης-ενέργειας (S_t, A_t) όπως ήταν στον Q-learning, αλλά η πιθανότητα επιλογής της ενέργειας a, δεδομένου της πολιτικής που ακολουθείται με την χρήση της ανα-

Αλγοριθμος 2.5: *SARSA*

Initialize $Q(s,a)$, for all $s \in S, \alpha \in A(s)$ and $Q(\textit{terminal-state}\cdot) = 0$
for (each episode) **do**:
    Initilize S
    Choose A from S using a policy
    repeat(for each step of episode):
        Take action A, observe R,S'
        Choose A' from S' using a policy
        $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',A') - Q(S,A)\big]$
        $S \leftarrow S'; A \leftarrow A'$
    until S is terminal
end for

μενόμενης τιμής. Ο κανόνας της ανανέωσης των τιμών αυτήν την φορά είναι:

$$\begin{aligned} Q(S_t,A_t) &\leftarrow Q(S_t,A_t) + \alpha[R_{t+1} + \gamma \mathbb{E}\big[\, Q(S_{t+1},A_{t+1}|S_{t+1}\big] \; - Q(S_t,A_t)\big] \\ &\leftarrow Q(S_t,A_t) + \alpha[R_{t+1} + \gamma \sum_{\alpha} \pi(a|S_{t+1}) Q(S_{t+1},\alpha) - Q(S_t,A_t)\big] \end{aligned}$$

Δεδομένης κάθε επόμενης κατάστασης S_{t+1}, αυτός ο αλγόριθμος λειτουργεί ντετερμινιστικά. Η παραπάνω διαδικασία είναι πιο πολύπλοκη από τον απλό αλγόριθμο $SARSA$ αλλά ελαττώνει την διακύμανση εξαιτίας της τυχαίας επιλογής των ενεργειών A_{t+1}.

Αλγοριθμος 2.6: *SARSA Expected*

Initialize $Q(s,a)$, for all $s \in S, \alpha \in A(s)$ and $Q(\textit{terminal-state}\cdot) = 0$
for (each episode) **do**:
    Initialize S
    repeat(for each step of episode):
        Choose A from S using a policy
        Take action A, observe R,S'
        $V_{s'} = \sum_a \pi(s',a) Q(s',a)$
        $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma V_{s'} - Q(S,A)\big]$
        $S \leftarrow S'$
    until S is terminal
end for

2.3.4.5 N-step SARSA

Μια παραλλαγή των προηγούμενων μεθόδων επιτυγχάνεται από μια ενδιάμεση ενημέρωση των τιμών. Πιο συγκεκριμένα για την κάθε πρόβλεψη η ανανέωση των τιμών μπορεί κάθε φορά να γίνεται μετά από μια συγκεκριμένη περίοδο, δηλαδή μετά από μια ακολουθία ενεργειών. Συμβολίζοντας με n τον αριθμό των μεταβάσεων από τις ενέργειες που θα ακολουθήσουν, τότε η μέθοδος αυτή ονομάζεται $n-step$ καθώς όλη η σύγκριση και πρόβλεψη γίνεται μετά από n βήματα. Όλοι οι παραπάνω αλγόριθμοι αποτελούν μια υποπερίπτωση της μεθόδου $n-step$ για $n=1$. Έτσι όταν η ενημέρωση των τιμών μετά τα n βήματα διεκπεραιώνεται με βάση την ανανέωση που πραγματοποιεί ο αλγόριθμος $SARSA$, προκύπτει η μέθοδος $n-step\ SARSA$.

Αλγοριθμος 2.7: *n-step SARSA for estimating* $Q \approx q_*$ *or* $Q \approx q_\pi$

Initialize $Q(s,a)$ for all $s \in S$, $\alpha \in S$
Initialize policy π
Parameters: step size $\alpha \in (0,1]$ and a positive integer n

Repeat(for each episode)
Initialize and store $S_0 \neq$ terminal
Select and store an action A_0 $\pi(\cdot|S_0)$
$T \leftarrow \infty$
for $t = 0, 1, 2 \ldots$ **do**:
    if $t < T$, **then**:
        Take action A_t
        Observe and store the next reward R_{t+1} and the next state as S_{t+1}
        if S_{t+1} is terminal **then**:
            $T \rightarrow t+1$
        else:
            Select and store an action A_{t+1} $\pi(\cdot|S_{t+1})$
        end if
    end if
    $\tau \leftarrow t-n+1$
    if $\tau \geq 0$ **then**:
        $G \leftarrow \sum_{i=\tau+1}^{min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
        if $\tau + n < T$ **then**:
            $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$
        end if
        $Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha[G - Q(S_\tau, A_\tau)]$
    end if
end for
$\tau = T - 1$

Κεφάλαιο 3

Περιγραφή θέματος

Στο κεφάλαιο αυτό παρουσιάζονται οι βασικές έννοιες για την περαιτέρω ανάλυση των Συστημάτων Αποθήκευσης Ενέργειας, των Αγορών Ενέργειας και την μεταξύ τους σύνδεση.

3.1 Συστήματα Αποθήκευσης Ενέργειας

3.1.1 Σημασία Αποθήκευσης Ηλεκτρικής Ενέργειας

Η διακύμανση της ισχύος που παράγεται από τις διατάξεις των Ανανεώσιμων Πηγών Ενέργειας είναι σημαντικά αισθητή σε ημερήσια, ωριαία και εποχιακή βάση λόγω της εξάρτησής της από τα καιρικά φαινόμενα, με αποτέλεσμα σε ορισμένες περιπτώσεις να υπάρχει έλλειψη της παραγόμενης ισχύος από ΑΠΕ ενώ σε άλλες να παρατηρείται πλεόνασμα. Συνεπώς, γίνεται όλο και πιο συχνό φαινόμενο η χρονική αναντιστοιχία της παραγωγής, με την κατανάλωση ενέργειας όσο αυξάνεται το μερίδιο συμμετοχής των ΑΠΕ στα δίκτυα. Σε αυτές τις περιπτώσεις, η ενέργεια ενδέχεται να μην διαθέσιμη όταν απαιτείται ή να υπάρχει σε αφθονία όταν η ζήτησή της είναι χαμηλή[15]. Όλη αυτή η αυξημένη μεταβλητότητα στα συστήματα ισχύος θέτει σε κίνδυνο την αξιόπιστη και ασφαλή λειτουργία του δικτύου[16][17].Συνεπώς η ευελιξία του δικτύου, δηλαδή ο βαθμός στον οποίο μπορεί το δίκτυο να διατηρήσει το ισοζύγιο ισχύος, δεδομένου των διακυμάνσεων, αποτελεί τον βασικό παράγοντα για να καταστεί δυνατή η περαιτέρω διείσδυση των ΑΠΕ.

Επομένως εξαιτίας αυτής της ασυμβατότητας της παραγωγής με την ζήτηση, η αποθήκευση της ηλεκτρικής ενέργειας που παράγεται από ΑΠΕ αποτελεί απαραίτητη προϋπόθεση τόσο για την περαιτέρω διείσδυσής τους στα δίκτυα όσο και για την ομαλή ενσωμάτωσή τους σε αυτά. Ακόμα η συνεχής μείωση του κόστους κατασκευής των συστημάτων αποθήκευσης ηλεκτρικής ενέργειας, σε συνδυασμό με την μεγάλη ευελιξία που παρέχουν στην λειτουργία του δικτύου, καθιστούν την αποθήκευση μια αρκετά αποδοτική και προσιτή λύση.[18]

Επομένως με την αποθήκευση ελαχιστοποιούνται οι απορρίψεις παραγωγής από ΑΠΕ, το οποίο εκτός από το γεγονός ότι ενισχύονται οι προσπάθειας για επίτευξη των στόχων απανθρακοποίησης μπορεί να εξασφαλίσει χαμηλού κόστους ηλεκτρική ενέργεια. Αξιοποιώντας αυτήν την ενέργεια τις ώρες αιχμής, όπου το κόστος της είναι υψηλότερο, μπορεί να επιτευχθεί οικονομική ελάφρυνση για τον τελικό καταναλωτή.

Η αποθήκευση ενέργειας αναμένεται να αυξήσει την ασφάλεια του συστήματος καθώς μπορεί να παρέχει αδιάκοπη λειτουργία στους καταναλωτές. Ακόμα ενισχύεται η ενεργειακή αυτονομία με αποτέλεσμα να βελτιώνεται η διαχείριση της ενέργειας. Για την κάθε περίπτω-

ση εφαρμογής του συστήματος αποθήκευσης η βέλτιστη αξιοποίησή του εξαρτάται από την απόδοση και την ποσότητα της ενέργειας που πρέπει κάθε φορά να διαχειριστεί.[19]

Επίσης τα συστήματα αποθήκευσης αξιοποιούνται ως μονάδες ταχείας εφεδρείας, για την άμεση εξυπηρέτηση της κατανάλωσης, σε περίπτωση απρόσμενης διακοπής λειτουργίας μίας εκ των μονάδων παραγωγής. Ακόμα μπορούν να έχουν εφαρμογή στον έλεγχο της συχνότητας και της μεταφοράς ισχύος εντός περιοχής του δικτύου.[20]

3.1.2 Τεχνολογίες Αποθήκευσης Ηλεκτρικής Ενέργειας

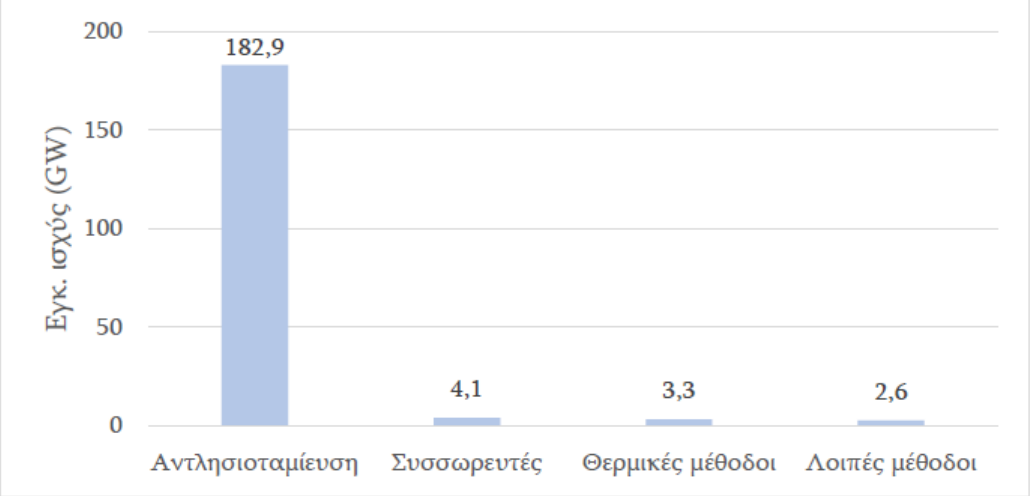
Τα Συστήματα Αποθήκευσης Ενέργειας (Energy Storage System - ESS) αποτελούν μια σύγχρονη τεχνολογία που έχει αναπτυχθεί για την αποτελεσματική αποθήκευση ηλεκτρικού φορτίου σε ειδικά διαμορφωμένα συστήματα. Η βασική ιδέα είναι να μπορέσει να χρησιμοποιηθεί η παραγόμενη ενέργεια μια άλλη χρονική στιγμή. Η αυξημένη ανάγκη για την μελλοντική εκμετάλλευση της παραγόμενης ενέργειας, έχει οδηγήσει σε μεγάλο ενδιαφέρον για την έρευνα και την ανάπτυξη των συστημάτων αποθήκευσης ενέργειας.

Υπάρχει μεγάλος αριθμός τεχνολογιών αποθήκευσης ηλεκτρικής ενέργειας με πολύ διαφορετικά τεχνικά ξαρακτηριστικά η κάθε μία. Γενικά, μπορούν να διακριθούν στις εξής κατηγορίες[21]:

- Ηλεκτροχημικές τεχνολογίες αποθήκευσης (συσσωρευτές)
- Μηχανικές τεχνολογίες αποθήκευσης (π.χ. συστήματα αντλησιοταμίευσης, συμπιεσμένου αέρα)
- Θερμικές τεχνολογίες αποθήκευσης (π.χ. αποθήκευση με υγροποίηση αέρα)
- Ηλεκτρομαγνητικές τεχνολογίες αποθήκευσης (π.χ. υπερπυκνωτές)

Όπως φαίνεται από το διάγραμμα 3.1 κυρίαρχη τεχνολογία αποθήκευσης ηλεκτρικής ενέργειας μέχρι σήμερα είναι με διαφορά η αντλησιοταμίευση. Για το λόγο αυτό, στη συνέχεια αναλύεται με περισσότερες λεπτομέρειες.

[1]Πηγή: Εισήγηση της ΟΔΕ - ΥΠΕΝ, 2021



Σχήμα 3.1: *Εγκατεστημένη ισχύς ανά τεχνολογία αποθήκευσης για το έτος 2020* [1]

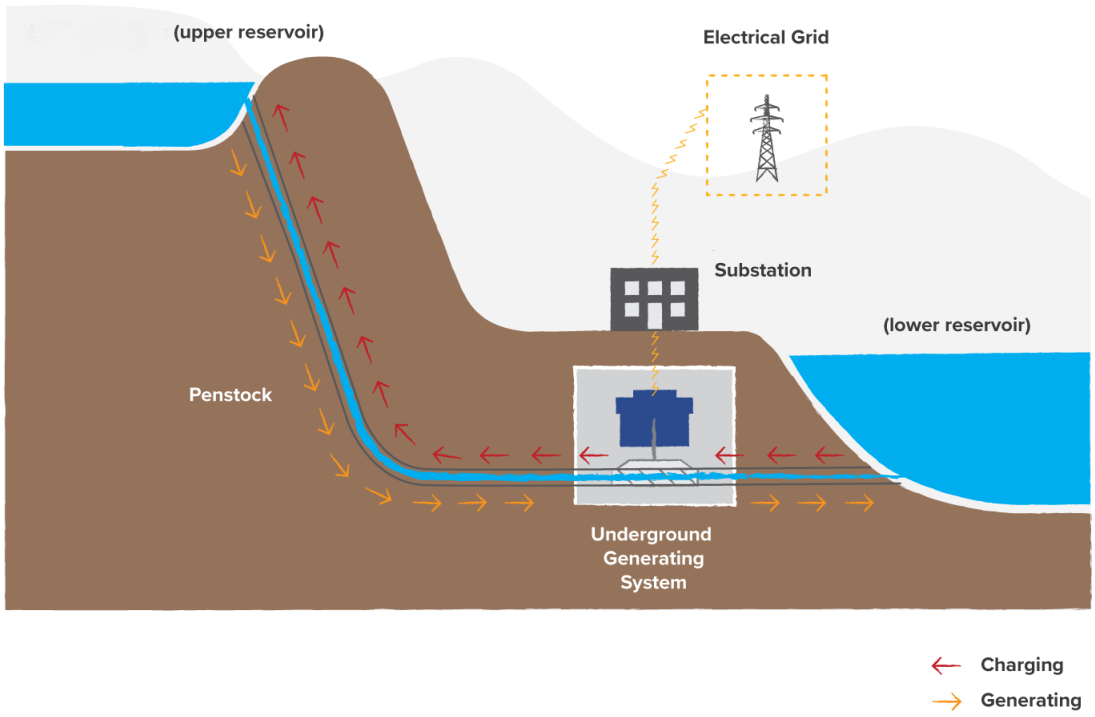
3.1.3 Σύστημα Αποθήκευσης με Αντλησιοταμίευση (Pumped Hydro Storage System)

Η άντληση των υδάτων με σκοπό την αποθήκευση ενέργειας αποτελεί μια από τις διαδεδομένες και παλαιότερες τεχνολογίες αποθήκευσης για ηλεκτροπαραγωγή μεγάλης κλίμακας. Ξεκίνησε να εμφανίζεται από το 1890 και πλέον αποτελεί παγκοσμίως την πιο ανεπτυγμένη και οικονομικά συμφέρουσα επιλογή για εφαρμογές αποθεματικού ενέργειας μεγάλης κλίμακας. Η δεδομένη εγκατεστημένη ισχύς συστημάτων αντλησιοταμίευσης φτάνει τα 181 GW, ποσό που αντιστοιχεί περίπου στο 95% των συστημάτων αποθήκευσης όλων των διαθέσιμων τεχνολογιών που βρίσκονται σε λειτουργία παγκοσμίως, ενώ η αποθηκευτική χωρητικότητα ξεπερνά τις 1.6 TWh [22]. Υπολογίζεται ότι μέχρι το 2050, η εγκατεστημένη ισχύς θα έχει φτάσει τα 325 GW [23].

Η αντλησιοταμίευση αποτελεί την πιο ελκυστική μέθοδο αποθήκευσης μεγάλης κλίμακας και μπορεί να χρησιμοποιηθεί σε διασυνδεδεμένα αλλά και σε αυτόνομα συστήματα. Η αρχή λειτουργίας της στηρίζεται στην δυναμική ενέργεια μεταξύ δύο ταμιευτήρων νερού που είναι κατασκευασμένοι σε υψομετρική διαφορά εκατοντάδων μέτρων. Ο κύριος εξοπλισμός ενός συστήματος αντλησιοταμίευσης (pumped hydro storage system) αποτελείται από αντλίες, γεννήτριες, αγωγούς και ένα σύστημα ελέγχου για την σύνδεση και την σωστή λειτουργία των δύο ταμιευτήρων νερού. Η σύνδεση του άνω και του κάτω ταμιευτήρα πραγματοποιείται με έναν ή με δύο αγωγούς πτώσης. Το πλεόνασμα της παραγόμενης ηλεκτρικής ενέργειας από ΑΠΕ σε περιόδους χαμηλής ζήτησης τροφοδοτεί ένα σύστημα αντλιών, μέσω των οποίων ανυψώνεται διά των αγωγών από τον κάτω ταμιευτήρα προς τον άνω, δίνοντας έτσι τη δυνατότητα αποθήκευσης της περίσσειας ηλεκτρικής ενέργειας με τη μορφή δυναμικής ενέργειας. Αντίστοιχα σε περιόδους αιχμής απελευθερώνεται το νερό του άνω ταμιευτήρα μετατρέποντας την δυναμική ενέργεια σε ηλεκτρική μέσω γεννητριών[24]. Η παραπάνω διαδικασία απεικονίζεται και στο σχήμα 3.2.

Το πιο σημαντικό πλεονέκτημα της μεθόδου αποτελεί η πολύ γρήγορη ανταπόκριση του συ-

[2]Πηγή: https://leapshydro.com



Σχήμα 3.2: *Σύστημα Αντλησιοταμίευσης*[2]

στήματος τόσο κατά την εκκίνηση όσο και κατά την διακοπή λειτουργίας καθώς ολοκληρώνεται εντός δευτερολέπτων, χαρακτηριστικό που προσδίδει μεγάλη ευελιξία στην ορθή λειτουργία του υπόλοιπου δικτύου. Επίσης ένα έργο αντλησιοταμίευσης έχει πολύ μεγάλη διάρκεια ζωής, με αποτέλεσμα η αξιοποίησή του να συνεχίζεται σε βάθος χρόνου. Το πιο βασικό ελάττωμα ανάπτυξης της μεθόδου είναι το πολύ υψηλό κόστος κατασκευής των τεχνητών ταμιευτήρων, συμπεριλαμβανομένων των γεωλογικών, γεωγραφικών και περιβαλλοντικών περιορισμών.

3.2 Αγορές Ενέργειας

Η παραδοσιακή λειτουργία της βιομηχανίας ηλεκτρικής ενέργειας αποτελούσε μια πλήρως καθετοποιημένη διαδικασία, υπό τον έλεγχο μιας και μόνο δημόσιας επιχείρησης, η οποία ήταν υπεύθυνη για το σύνολο των δραστηριοτήτων στην ηλεκτρική ενέργεια. Οι δραστηριότητες αυτές αποτελούνται από: την παραγωγή, την μεταφορά, την διανομή και την προμήθεια της ηλεκτρικής ενέργειας. Κατά τη διάρκεια της δεκαετίας του 1990 στις περισσότερες χώρες της Ευρωπαϊκής Ένωσης, μια από τις βασικές μεταρρυθμίσεις στον τομέα της ενέργειας, ήταν η απελευθέρωση της βιομηχανίας της ηλεκτρικής ενέργειας και η δημιουργία μιας ενιαίας ανταγωνιστικής αγοράς. Έτσι οι βασικές δραστηριότητες διαχωρίστηκαν και ελέγχονται από διαφορετικούς φορείς ενώ η προμήθεια ηλεκτρικής ενέργειας παρακολουθείται από τις αντίστοιχες ρυθμιστικές αρχές. Στο πλαίσιο αυτό, αναπτυχθηκαν χονδρικές αγορές ηλεκτρικής ενέργειας κατά μήκος της Ευρώπης, προκειμένου να εξυπηρετήσουν την ανταγωνιστική λειτουργία της βιομηχανίας ηλεκτρισμού και να προσφέρουν οικονομικές παροχές στους τελικούς καταναλωτές. Ανεξάρτητα από τη δομή και οργάνωση της απελευθερωμένης χονδρεμπορικής αγοράς κάθε Ευρωπαϊκής χώρας, ενέργεια και εφεδρείες αποτελούν προϊόν διαπραγμάτευσης, από το μακροπρόθεσμο επίπεδο έως και τον πραγματικό χρόνο. Στις σύγχρονες απελευθερωμένες χονδρεμπορικές αγορές ηλεκτρικής ενέργειας, οι συναλλαγές μεταξύ των συμμετεχόντων πραγματοποιούνται τόσο σε μακροπρόθεσμο επίπεδο όσο και σε πραγματικό χρόνο μέσω των αντίστοιχων αγορών.[25] Οι πιο κοινές κατηγορίες συναλλαγών που εμφανίζονται στις περισσότερες χώρες είναι η:

- Προθεσμιακή Αγορά (Forward Market)
- Η Προημερήσια Αγορά (Day-Ahead Market)
- Η Ενδοημερήσια Αγορά (Intraday Market)
- Η Αγορά Εξισορρόπησης ή Πραγματικού Χρόνου (Balancing or Real-Time Market)

Στην Ελλάδα για την λειτουργία των τριών πρώτων αγορών είναι υπεύθυνο το Ελληνικό Χρηματιστήριο Ενέργειας (ΕΧΕ), ενώ για τον συντονισμό της Αγορά Εξισορρόπησης, ο Ανεξάρτητος Διαχειριστής Μεταφοράς Ηλεκτρικής Ενέργειας (ΑΔΜΗΕ).

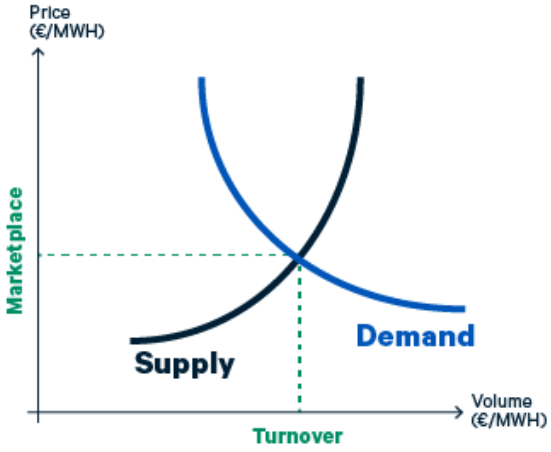
Στη συνέχεια, αναλύονται συνοπτικά τα βασικά χαρακτηριστικά των παραπάνω αγορών ηλεκτρικής ενέργειας[26].

3.2.1 Προθεσμιακή Αγορά (Forward Market)

Στην Προθεσμιακή Αγορά, πραγματοποιούνται συμφωνίες μεταξύ αγοραστών και πωλητών για την αγοραπωλησία ηλεκτρικής ενέργειας με στόχο την φυσική παράδοσή της σε μελλοντικό χρόνο και τον καθορισμό προσυμφωνημένες τιμών στο παρόν. Προθεσμιακές συναλλαγές μπορούν να πραγματοποιηθούν με διμερή έξω-χρηματιστηριακά συμβόλαια (bilateral Over-the-Counter Markets), ή μέσω του χρηματιστηρίου ενέργειας με συμβόλαια μελλοντικής εκπλήρωσης (Future Market) όπου οι συμμετέχοντες ανταλλάσσουν παράγωγα προϊόντα ηλεκτρικής ενέργειας υποχρεωτικής φυσικής παράδοσης (physically settled products) ή καθαρά οικονομικά (financial products) με στόχο είτε την αντιστάθμιση του ρίσκου έναντι πιθανών διακυμάνσεων της τιμής είτε το κέρδος. [27]

3.2.2 Προημερήσια Αγορά (Day Ahead Market)

Είναι η βασική αγορά, όπου αγοράζονται και πωλούνται οι ποσότητες ηλεκτρικής ενέργειας που θα παραχθούν και θα παραδοθούν την επόμενη μέρα. Η Προημερήσια Αγορά αναφέρεται στην χονδρεμπορική αγοραπωλησία ηλεκτρικής ενέργειας μεταξύ παραγωγών και προμηθευτών. Πρόκειται για μια δημοπρασία με δέσμευση φυσικής παράδοσης την ημέρα D και κλείσιμο των προσφορών από την ημέρα $D-1$. Με αυτόν τον τρόπο παράγει μια ενιαία τιμή εκκαθάρισης για κάθε περίοδο κατανομής. Η συμμετοχή είναι υποχρεωτική μόνο για τους παραγωγούς και προαιρετική για όλους τους άλλους συμμετέχοντες. Η συναλλαγή είναι εφικτή είτε με διμερή συμβόλαια OTC (Over The Counter), είτε με ανταλλαγή ισχύος. Η Αγορά Day-Ahead λειτουργεί σε πραγματικό χρόνο. Η τιμή της κάθε ώρας της επόμενης ημέρας εξαρτάται από τις προσφορές έγχυσης και τις δηλώσεις φορτίου που υποβάλλονται. Οι συμμετέχοντες χρεώνονται ή πιστώνονται την οριακή τιμή του συστήματος.



Σχήμα 3.3: *Προσφορά-Ζήτηση στην Προημερήσια Αγορά*[3]

Το σημείο τομής των καμπυλών προσφοράς και ζήτησης αποτελεί την οριακή τιμή.

3.2.3 Ενδο-ημερήσια Αγορά (Intraday Market)

Η ενδο-ημερήσια αγορά πραγματοποιείται κατά την ημέρα φυσικής παράδοσης D, όπου δίνει την ευκαιρία στους συμμετέχοντες να βελτιώνουν την θέση τους κατά την διάρκεια της ημέρας εφόσον έχουν προκύψει αποκλίσεις από τις προσφορές τους. Οι συμμετέχοντες έχοντας στην διάθεσή τους νέες πληροφορίες, αλλά και γνωρίζοντας τα αποτελέσματα από την εκκαθάριση της προηγούμενης μέρας μπορούν να ελαχιστοποιήσουν περισσότερο τις αποκλίσεις των θέσεών τους. Στην Ενδοημερήσια Αγορά, η συμμετοχή είναι προαιρετική για όλους τους συμμετέχοντες.

3.2.4 Αγορά Εξισορρόπησης (Balancing Market)

Η Αγορά Εξισορρόπησης αποτελεί αγορά πραγματικού χρόνου, με σκοπό την διόρθωση της ανισορροπίας μεταξύ παραγωγής και ζήτησης από τον Διαχειριστή του Συστήματος. Έτσι

[3]Πηγή: https://www.nordpoolgroup.com/the-power-market/Day-ahead-market/

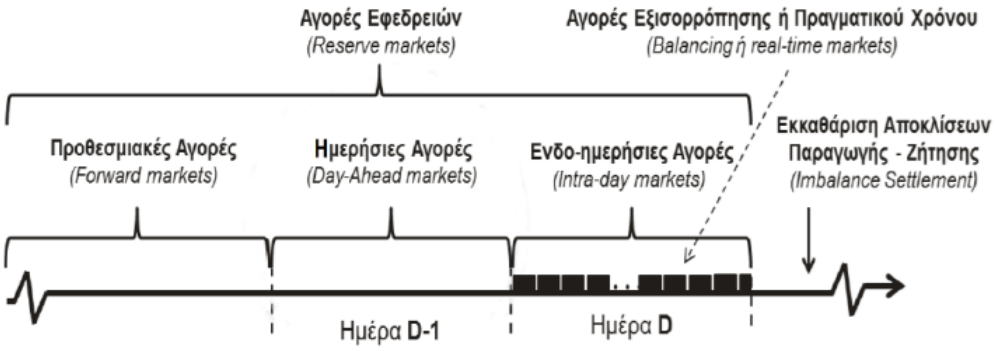
ενεργοποιώντας ανοδικές ή καθοδικές προσφορές εξισορρόπησης, για αρνητικές ή θετικές αποκλίσεις του συστήματος αντίστοιχα, πετυχαίνει την αποκατάσταση της ισορροπίας του. Οι συμμετέχοντες πληρώνονται από τον Διαχειριστή για ενδεχόμενη αύξηση της παραγωγής τους ή μείωση της κατανάλωσής τους, ενώ χρεώνονται για μείωση της παραγωγής τους ή αύξηση της κατανάλωσής τους. Η αγορά εξισορρόπησης αποτελείται από τρεις επιμέρους αγορές:

- την Αγορά Ισχύος Εξισορρόπησης. Στόχος της αγοράς είναι να εξασφαλίζει την επάρκεια ισχύος για την κάλυψη των αναγκών του δικτύου. Σε αυτό το στάδιο οι συμμετέχοντες υποβάλλουν προσφορές για τις απαιτούμενες εφεδρείες που μπορούν να προσφέρουν και είναι αναγκαίες στο σύστημα. Αντίστοιχα αποζημιώνονται βάσει της τιμής προσφοράς (pay-as-bid). Στον σχεδιασμό της αγοράς, έχουν προδιαγραφεί τρία βασικά προϊόντα ισχύος εξισορρόπησης:

  1. Η ανοδική και καθοδική Εφεδρεία Διατήρησης Συχνότητας (Frequency Containment Reserve, FCR)
  2. Η ανοδική και καθοδική χειροκίνητη Εφεδρεία Αποκατάστασης Συχνότητας (manual Frequency Restoration Reserve – mFRR)
  3. Η ανοδική και καθοδική αυτόματη Εφεδρεία Αποκατάστασης Συχνότητας (automatic Frequency Restoration Reserve – aFRR).

- την Αγορά Ενέργειας Εξισορρόπησης. Αποτελεί την αγορά πραγματικού χρόνου, στην οποία πραγματοποιείται ενεργοποίηση των προϊόντων ενέργειας εξισορρόπησης βάσει τιμών προσφορών και βάσει των αναγκών του Συστήματος.

- την Εκκαθάριση Αποκλίσεων. Πραγματοποιείται μετά τον πραγματικό χρόνο και ολοκληρώνεται ο επιμερισμός του κόστους σε όλους τους συμμετέχοντες που προκάλεσαν αποκλίσεις σε πραγματικό χρόνο.[28]



Σχήμα 3.4: *Αναπαράσταση χρονισμού των επιμέρους αγορών*[4]

[4]Πηγή: Περιοδικό Ένεργών, Τεύχος 6, ΑΔΜΗΕ

3.3 Συμμετοχή της αποθήκευσης στην αγορά

Ένα Σύστημα Αποθήκευσης Ηλεκτρικής Ενέργειας μπορεί να συμμετέχει στις αγορές ενέργειας (προθεσμιακή, προημερήσια, ενδο-ημερήσια) μέσω arbitrage, δηλαδή να πραγματοποιεί φόρτιση σε ώρες χαμηλής ζήτησης ή υψηλής παραγωγής των ΑΠΕ με σκοπό την εκφόρτισή του σε ώρες υψηλής ζήτησης ή μειωμένης παραγωγής ΑΠΕ, καθώς και με σκοπό τη συμμετοχή στην αγορά εξισορρόπησης ισχύος και ενέργειας. Στα πλαίσια της εργασίας μελετάται η συμμετοχή του Συστήματος Αποθήκευσης σε μια αγορά εξισορρόπησης.

Με την απελευθέρωση της αγοράς ενέργειας, οι Διαχειριστές Συστήματος Μεταφοράς δεν έχουν στην κατοχή τους μονάδες παραγωγής. Για την εγγύηση της αξιοπιστίας και της σταθερότητας του συστήματος, ο ρόλος της εξισορρόπησης της παραγωγής ηλεκτρικής ενέργειας με την κατανάλωση, υποστηρίζεται από συμμετέχοντες στην αγορά, γνωστοί ως φορείς ευθύνης εξισορρόπησης (Balance Responsible Parties-BRPs). Για να επιτευχθεί αυτό, έχουν δημιουργηθεί αγορές εφεδρειών οι οποίες συναλλάσσονται διαδοχικά με τις αγορές ενέργειας (Προθεσμιακή, Προημερήσια, Ενδοημερήσια, Εξισορρόπησης), μέσω ανεξάρτητων δημοπρασιών.

Με βάση αυτό το σύστημα, οι BRPs είναι υπεύθυνοι για την συνεχή εξισορρόπηση του φορτίου τους με την παραγωγή, λαμβάνοντας υπόψη την δραστηριότητα όλων των υπόλοιπων BRPs που συμμετέχουν στην αγορά. Έτσι, η υπολειπόμενη ανισορροπία στο σύστημα, αντισταθμίζεται σε πραγματικό χρόνο από τον Διαχειριστή, χρησιμοποιώντας τις διαθέσιμες επικουρικές υπηρεσίες από την αγορά εφεδρειών.

Το κόστος που προκύπτει από την ενεργοποίηση των υπηρεσιών καλύπτεται χρεώνοντας κάθε φορέα BRP που βρίσκεται σε ανισορροπία. Οι χρεώσεις αυτές υπολογίζονται σε βάση τετάρτου της ώρας και εφαρμόζονται κατά την διάρκεια που πραγματοποιείται η ανισορροπία στο σύστημα. Επομένως, ο μηχανισμός αυτός αποτελεί την εκκαθάριση αποκλίσεων της παραγωγής με την ζήτηση (Imbalance settlement). Η εκκαθάριση των αποκλίσεων θα πραγματοποιείται σύμφωνα με την αρχή "single price", δηλαδή σε ενιαία τιμή ανεξάρτητα από την κατεύθυνση που θα έχει κάθε φορά η απόκλιση και θα ισούται με την μεσοσταθμική τιμή όλων των ενεργοποιημένων προσφορών ενέργειας εξισορρόπησης εντός της περιόδου εκκαθάρισης. Αυτός ο μηχανισμός ευνοεί τους BRPs που βοηθούν στην αποκατάσταση της ανισορροπίας του συστήματος (πράσινα πεδία στον πίνακα 3.5), ενώ επιβάλλει ποινές σε όσους επιδεινώνουν την ανισορροπία (κόκκινα πεδία στον πίνακα 3.5).

Η τιμή ανισορροπίας του συστήματος εξαρτάται από την ποσότητα των εφεδρειών που θα χρησιμοποιηθούν από τον Διαχειριστή. Όταν υπάρχει έλλειψη παραγόμενης ενέργειας στο δίκτυο, ο διαχειριστής πρέπει να ενεργοποιήσει τον μηχανισμό ανοδικής Εφεδρείας, το κόστος του οποίου αποτελεί την οριακή τιμή ανοδικής ρύθμισης (Marginal Incremental Price - MIP). Οι BRPs που είναι υπεύθυνοι για αυτήν την έλλειψη, είναι υποχρεωμένοι να πληρώσουν αυτό το κόστος (το οποίο συνήθως είναι υψηλότερο από το κόστος της ανάλογης ενέργειας στην προημερήσια αγορά), ενώ για τους φορείς με πλεόνασμα παραγωγής η οριακή τιμή ανοδικής ρύθμισης εκφράζεται με την μορφή κέρδους. Αντίστοιχα, όταν υπάρχει πλεόνασμα παραγωγής στο σύστημα, ενεργοποιείται ο μηχανισμός καθοδικής Εφεδρείας, η τιμή του οποίου ορίζεται από την οριακή τιμή καθοδικής ρύθμισης (Marginal Decremental Price - MDP). Οι BRPs με πλεόνασμα ενέργειας παραγωγής κερδίζουν σε αυτή την περίπτωση την οριακή τιμή (η

οποία όμως, συνήθως είναι χαμηλότερη από την ανάλογη τιμή ενέργειας στην προημερήσια αγορά) για το πλεόνασμά τους. Αντιθέτως οι BRPs που βρίσκονται σε έλλειψη παραγωγής, θα επιβαρυνθούν με χαμηλό κόστος για την αντίστοιχη αρνητική ανισορροπία τους, καθώς βοηθούν στην αποκατάσταση της ισορροπίας του συστήματος.

Η τιμή ανισορροπίας ενέργειας, καθορίζει από την πλευρά των φορέων εξισορρόπησης, την τιμή της ενέργειας που χρησιμοποιήθηκε για να καλύψει το δίκτυο, το ισοζύγιο ισχύος. Σε ένα ιδανικά ισορροπημένο δίκτυο, χωρίς καμία απόκλιση στις προβλέψεις, επομένως χωρίς ανάγκη για κάποια εξισορρόπηση, η τιμή αυτή θα ήταν 0 €/MWh. Εξαιτίας της μεγάλης διακύμανσης των τιμών MIP και MDP, επικρατεί μεγάλη αστάθεια στην τιμή της ανισορροπίας. Ωστόσο με την αξιοποίηση της αποθήκευσης, μπορεί να ελαττωθεί αυτή η υψηλή διακύμανση και ταυτόχρονα να εξασφαλιστεί οικονομικό όφελος για τον κάθε φορέα που συμμετέχει στην αγορά και τον τελικό καταναλωτή.[29]

		Διαχειριστής Μεταφοράς - TSO	
		Πλεόνασμα Θετική ανισορροπία Παραγωγή > Ζήτηση	*Έλλειψη* Αρνητική ανισορροπία Παραγωγή < Ζήτηση
BRPs	*Πλεόνασμα*	Ο Διαχειριστής πληρώνει τους BRPs	Ο Διαχειριστής πληρώνει τους BRPs
	Έλλειψη	Οι BRPs πληρώνουν τον Διαχειριστή	Οι BRPs πληρώνουν τον Διαχειριστή

Σχήμα 3.5: *Πίνακας εκκαθάρισης αποκλίσεων*

Κεφάλαιο 4

Ανάλυση και σχεδίαση

4.1 Περιγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν περιγράφουν την ανισορροπία του δικτύου της Γερμανίας στην αγορά ενέργειάς της τα έτη 2015-2016. Πιο συγκεκριμένα, πρόκειται για μια χρονοσειρά χωρισμένη σε διαστήματα 15 λεπτών της ώρας που αποτελείται από τις τιμές ανισορροπίας ενέργειας (Imbalance Price) ($€/MWh$) και την ποσότητα της ισχύς ανισορροπίας (Imbalance Volume) (MW). Για τις ανάγκες της διπλωματικής χρησιμοποιήθηκαν μόνο οι τιμές ανισορροπίας.

4.2 Μοντέλο Αποθήκευσης

Η σχεδίαση του μοντέλου του συστήματος αποθήκευσης αν και αποτελεί μια πολύπλοκη και σύνθετη διαδικασία είναι καθοριστικός παράγοντας για την σωστή προσομοίωση του πειράματος. Για τις ανάγκες της διπλωματικής αρχικά παραμετροποιήθηκε η ποσότητα ενέργειας που είναι εφικτό να αποθηκευτεί στο σύστημα ή να διοχετευθεί στο δίκτυο εντός ενός χρονικού διαστήματος προσομοίωσης dt, σύμφωνα με τον παρακάτω κανόνα[30][31]:

$$S_{change} = [\eta^{ch} \times P^{ch} - \frac{P^{dis}}{\eta^{dis}}] \times dt \tag{4.1}$$

όπου:

η^{ch} (charge efficiency) : Η απόδοση φόρτισης ισχύος του συστήματος αποθήκευσης

η^{dis} (discharge efficiency) : Η απόδοση εκφόρτισης ισχύος του συστήματος αποθήκευσης

P^{ch} (charge power) : Η ποσότητα ισχύος (MW) που φορτίστηκε στο σύστημα την δεδομένη χρονική στιγμή. Η ποσότητα αυτή περιορίζεται σε μια μέγιστη τιμή φόρτισης $\overline{P}$ όπου, $0 \leq P^{ch} \leq \overline{P}$

P^{dis} (discharge power) : Η ποσότητα ισχύος (MW) που εκφορτίστηκε από το σύστημα την δεδομένη χρονική στιγμή και παραχωρήθηκε στο δίκτυο. Η ποσότητα αυτή περιορίζεται αντίστοιχα σε μια μέγιστη τιμή εκφόρτισης $\underline{P}$, δηλαδή $0 \leq P^{dis} \leq \underline{P}$

dt (interval time) : το χρονικό διάστημα που πραγματοποιείται η προσομοίωση, κατέπέκταση ο χρόνος που απαιτείται για την απόκριση του συστήματος αποθήκευσης.

4.3 Σχεδιασμός και υλοποίηση του Συστήματος

Όλα τα πειράματα εκτελέστηκαν με βάση τα πρότυπα του πακέτου Gym της OpenAI [32]. Μέσω του μοντέλου και των δυνατοτήτων του Gym γίνεται ευκολότερα όλη η διαδικασία του σχεδιασμού, της υλοποίησης και της εκτίμησης των διαφορετικών αλγορίθμων για το δεδομένο περιβάλλον.

4.3.1 Περιβάλλον

Για να εκτιμηθεί η απόδοση των αλγορίθμων είναι αναγκαία η πιστή προσομοίωση του συστήματος αποθήκευσης σε συνδυασμό με τις τιμές ανισορροπίας για να υπολογιστεί το τελικό κέρδος. Η σωστή κατασκευή του περιβάλλοντος προσομοίωσης πραγματοποιείται όταν αυτό υπακούει στους ίδιους κανόνες και προσεγγίζει τα ίδια αποτελέσματα που θα επέστρεφε ένα αντίστοιχο ρεαλιστικό μοντέλο όταν αυτό δεχτεί τις ίδιες ενέργειες. Χαρακτηριστικά του περιβάλλοντος του συστήματος είναι η δυνατότητα φόρτισης ή εκφόρτισης της επιθυμητής ποσότητας ισχύος οποιαδήποτε χρονική στιγμή για ένα διάστημα dt, καθώς και ο υπολογισμός του κέρδους που θα αποφέρει στον χειριστή η κάθε απόφασή του.

Λαμβάνοντας υπόψη τα πρότυπα του Gym, έγινε αρχικά η υλοποίηση των συναρτήσεων λειτουργίας του Περιβάλλοντος με τις κατάλληλες παραμέτρους για να προσομοιώνει ένα Σύστημα Αποθήκευσης με τις ανάλογες προδιαγραφές. Τα ορίσματα που δέχεται σαν είσοδο το Περιβάλλον είναι:

- Το αρχείο των δεδομένων του data set. Γίνεται εισαγωγή όλων των δεδομένων ανισορροπίας της αγοράς.
- Η αναλογία των δεδομένων εκπαίδευσης (train) και ελέγχου (test).
- Το χρονικό διάστημα dt (interval) που θα μεσολαβεί μεταξύ των ενεργειών και, συνεπώς, ο χρόνος που θα πραγματοποιείται η προσομοίωση.
- Τα τεχνικά χαρακτηριστικά των προδιαγραφών του συστήματος αποθηκευσης.

Για την λειτουργία του συστήματος, το πρόβλημα πρέπει να μοντελοποιηθει ως μια Μαρκοβιανή Διαδικασία Αποφάσεων. Έτσι κατασκευάζεται μια διαδικασία που λειτουργεί για κάθε χρονική στιγμή t.

4.3.2 Καταστάσεις (States)

Μετά από κάθε απόφαση που λαμβάνει ο Πράκτορας θα πρέπει να γίνεται και η αντίστοιχη διαδικασία από το Περιβάλλον για να πραγματοποιείται η μετάβαση στην επόμενη νέα κατάσταση. Για την εφαρμογή της έννοιας των καταστάσεων $S_t \in S$ ορίζεται αντίστοιχα ένας χώρος παρατηρήσεων (Observation space), η διάσταση του οποίου ισούται με τον αριθμό των καταστάσεων. Κάθε κατάσταση του χώρου χρησιμοποιήθηκε για να μπορέσει ο Πράκτορας να εκτιμήσει καλύτερα το Περιβάλλον και συνεπώς να πετύχει πιο εύκολα τον στόχο του κατά την εκπαίδευση.

Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκαν πέντε διαφορετικές καταστάσεις οι οποίες περιγράφονται από την πεντάδα :

$$(SoC_t,\ CP_t,\ DP_t,\ \lambda^{SI}_{t-1},\ D_t) \tag{4.2}$$

και αναλύονται στην συνέχεια:

1. SoC_t (State of Charge): Η πρώτη και πιο βασική κατάσταση του χώρου αποτελεί η κατάσταση φόρτισης του συστήματος αποθήκευσης, δηλαδή τα αποθέματα της ενέργειας (MWh) που είναι διαθέσιμα προς αξιοποίηση. Με την ένδειξη της SoC_t φαίνεται το επίπεδο φόρτισης του συστήματος σε σχέση με τη χωρητικότητά του την χρονική στιγμή t. Δεδομένου ότι η χωρητικότητα του συστήματος αποθήκευσης είναι $\overline{S}$, η SoC_t θα περιγράφεται από συνεχείς τιμές στο διάστημα $0 \leq SoC_t \leq \overline{S}$ και μετά από κάθε νέα φόρτιση ή εκφόρτιση θα ενημερώνεται σύμφωνα με τον κανόνα:

   $$SoC_{t+1} = SoC_t + S_{change}$$

   όπου S_{change} δίνεται από την εξίσωση (4.1).

2. CP_t (Charge Potential) : Η δεύτερη διάσταση αφορά την διαθέσιμη ποσότητα ισχύος (MW) που μπορεί να υποστηρίξει το σύστημα, ανάλογα με τα κατασκευαστικά χαρακτηριστικά του, για το χρονικό διάστημα που πραγματοποιείται η προσομοίωση. Έτσι η CP_t επιστρέφει την ποσότητα που είναι δυνατό να φορτιστεί την χρονική στιγμή t στο σύστημα αποθήκευσης. Όπως η τιμή της P^{ch} στην (4.1) περιορίζεται από την μέγιστη τιμή $\overline{P}$, έτσι και για την CP_t ισχύει αντίστοιχα ότι $0 \leq CP_t \leq \overline{P}$.

3. DP_t (Discharge Potential) : Αντίστοιχα η επόμενη κατάσταση περιγράφει την διαθέσιμη ποσότητα ισχύος (MW) που είναι δυνατό να εγχυθεί στο δίκτυο από το σύστημα την χρονική στιγμή t και ισχύει ότι $0 \leq DP_t \leq \underline{P}$.

4. λ^{SI}_{t-1} (Imbalance Price) : Η τέταρτη κατάσταση του χώρου επιστρέφει κάθε φορά την τιμή ανισορροπίας $(€/MWh)$ για το σύστημα ανισορροπίας (System Imbalance) της αγοράς, την αμέσως προηγούμενη χρονική στιγμή από αυτήν που καλείται ο Πράκτορας να πάρει την απόφασή του.

5. D_t (Deviation) : Η τελευταία διάσταση εκφράζει την απόκλιση της τιμής ανισορροπίας $(€/MWh)$ της αμέσως προηγούμενης χρονικής στιγμής, από τον μέσο όρο των τελευταίων n τιμών,

   $$D_t = \frac{\sum^n \lambda^{SI}}{n} - \lambda^{SI}_{t-1}$$

4.3.3 Ενέργειες (Actions)

Παρομοίως λειτουργεί και ο χώρος των δυνατών ενεργείων (action space) από τον οποίο μπορεί να επιλέξει ο Πράκτορας μια επιθυμητή ενέργεια. Πρόκειται για έναν μονοδιάστατο χώρο, μέσω του οποίου, το Περιβάλλον δίνει την δυνατότητα στον Πράκτορα να επιλέξει μια τιμή για την ενέργεια $A_t \in A$ κάθε χρονικής στιγμής t. Η τιμή αυτή εκφράζει την φόρτιση

ή εκφόρτισης που θα πραγματοποιηθεί στο σύστημα αποθήκευσης. Οι τιμές που παίρνει είναι συνεχείς στο $A = [-1, 1]$ και εκφράζουν το ποσοστό επί τοις εκατό της μέγιστης δυνατής ισχύς που μπορεί να διαχειριστεί για το χρονικό διάστημα dt το σύστημα. Ισχύει ότι:

- για $A_t < 0$ πραγματοποιείται εκφόρτιση ενέργειας από το σύστημα αποθήκευσης και εγχέεται στο δίκτυο. Για $A_t = -1$ διοχετεύεται στο δίκτυο η μέγιστη δυνατή ενέργεια που μπορεί να παραχωρήσει στο δεδομένο χρονικό διάστημα dt το σύστημα αποθήκευσης.
- η τιμή $A_t = 0$ εκφράζει την αδράνεια του Πράκτορα, καθώς δεν δημιουργεί καμία μεταβολή στο σύστημα.
- ενώ για $A_t > 0$ πραγματοποιείται φόρτιση στο σύστημα αποθήκευσης από το δίκτυο. Για $A_t = 1$ αποθηκεύεται στο σύστημα η μέγιστη δυνατή ενέργεια που μπορεί να υποστηρίξει για το δεδομένο χρονικό διάστημα dt.

4.3.4 Ανταμοιβές (Rewards)

Η ανταμοιβή που επιστρέφεται μετά από κάθε επεισόδιο στον Πράκτορα, απαιτεί προσεκτικό σχεδιασμό καθώς αποτελεί τον καθοριστικό παράγοντα, τόσο για την ανάπτυξη της πολιτικής του, όσο και για την σωστή εκτίμηση της απόδοσής του. Αποτελεί ουσιαστικά, τον καθοριστικό παράγοντα για την ορθή λειτουργία όλων των μεθόδων της Ενισχυτικής Μάθησης και συχνά ο ορισμός της συνάρτησης που θα υπολογίζει την ανταμοιβή αυτή, καταλήγει να είναι μια ιδιαίτερα δύσκολη διαδικασία.

Το περιβάλλον του Συστήματος Αποθήκευσης πρέπει να επιστρέφει στον Πράκτορα τα οικονομικά αποτελέσματα της κάθε απόφασής του. Αρχικά το κέρδος ή η ζημία του Πράκτορα, υπολογίζεται από την νέα τιμή της ανισορροπίας που υπάρχει στην αγορά και της ποσότητας ενέργειας που επέλεξε να χρησιμοποιήσει. Για παράδειγμα όταν ο Πράκτορας φορτίζει το σύστημα για αρνητικές τιμές ανισορροπίας έχει κέρδος, και ζημία για θετικές. Για να μπορέσει το αποτέλεσμα να είναι ακόμα πιο ρεαλιστικό είναι σημαντικό να συνυπολογιστεί το κόστος από τις προμήθειες των συναλλαγών που μπορεί να επιβάλλει ένας aggregator, μέσω του οποίου γίνονται οι συναλλαγές στην αγορά εξισορρόπησης. Λαμβάνοντας υπόψη τις (4.2) & (4.1), η ανταμοιβή που θα λαμβάνει ο Πράκτορας θα δίνεται από την σχέση :

$$R_{t+1} = A_t \times |S_{change}| \times \lambda_t^{SI} - c_t \tag{4.3}$$

όπου c_t, είναι η χρέωση για την κάθε συναλλαγή που πραγματοποιεί ο Πράκτορας την χρονική στιγμή t.

Κεφάλαιο 5

Υλοποίηση

Στο κεφάλαιο αυτό παρουσιάζεται ο τρόπος υλοποίησης του συστήματος αποθήκευσης καθώς και όλα τα πειράματα που πραγματοποιήθηκαν, με βάση όσα αναφέρθηκαν στα προηγούμενα κεφάλαια.

5.1 Προσομοίωση

Αρχικά κάθε υλοποίηση αλγορίθμου πραγματοποιήθηκε με βάση τις ίδιες τεχνικές προδιαγραφές του συστήματος αποθήκευσης για να μπορέσει να γίνει με επιτυχία η εκτίμησή του έναντι των υπολοίπων. Στον παρακάτω πίνακα φαίνονται τα τεχνικά χαρακτηριστικά του συστήματος αποθήκευσης που χρησιμοποιήθηκαν για όλα πειράματα.

Energy Storage System	
Capacity $\overline{S}$	200 MWh
Max Charge Rate $\overline{P}$	200 MW
Max Discharge Rate $\underline{P}$	200 MW
Initial SoC	100 MWh
Minimum SoC	0 MWh
Charge Efficiency η^{ch}	75%
Discharge Efficiency η^{dis}	75%

Πίνακας 5.1: *Τεχνικά χαρακτηριστικά Συστήματος Αποθήκευσης*

Επίσης για τον συνυπολογισμό των προμηθειών της κάθε συναλλαγής ορίστηκε ενδεικτικά το ποσό που αντιστοιχεί στο 3% της αξίας της ενέργειας που συναλλάσσεται για το χρονικό διάστημα προσομοίωσης dt. Έτσι ο Πράκτορας το λαμβάνει σαν έξοδο και είναι πιθανό να μην τον συμφέρει να πραγματοποιεί ενέργειες σε κάθε χρονικό διάστημα. Είναι σημαντικό λοιπόν, να του δίνεται η δυνατότητα να μένει αδρανής, το οποίο το πετυχαίνει για μηδενική τιμή της ενέργειάς του και η προσομοίωση να πραγματοποιείται για το ελάχιστο δυνατό διάστημα με σκοπό να εκμεταλλεύεται όλη την διαθέσιμη πληροφορία από τα δεδομένα. Για να επιτευχθεί αυτό, χρησιμοποιήθηκε $dt = 15$ λεπτά για όλα τα πειράματα, δηλαδή το ίδιο χρονικό διάστημα ανα το οποίο είναι διαθέσιμες οι τιμές ανισορροπίας στα δεδομένα. Με αυτόν τον τρόπο θα έχει την δυνατότητα να πραγματοποιεί συναλλαγές εντός μεγαλύτερων χρονικών διαστημάτων εφόσον κρίνει ότι είναι ασύμφορη η συναλλαγή για μικρότερα χρονικά διαστήματα.

5.1.1 Α΄ Μέρος πειράματος

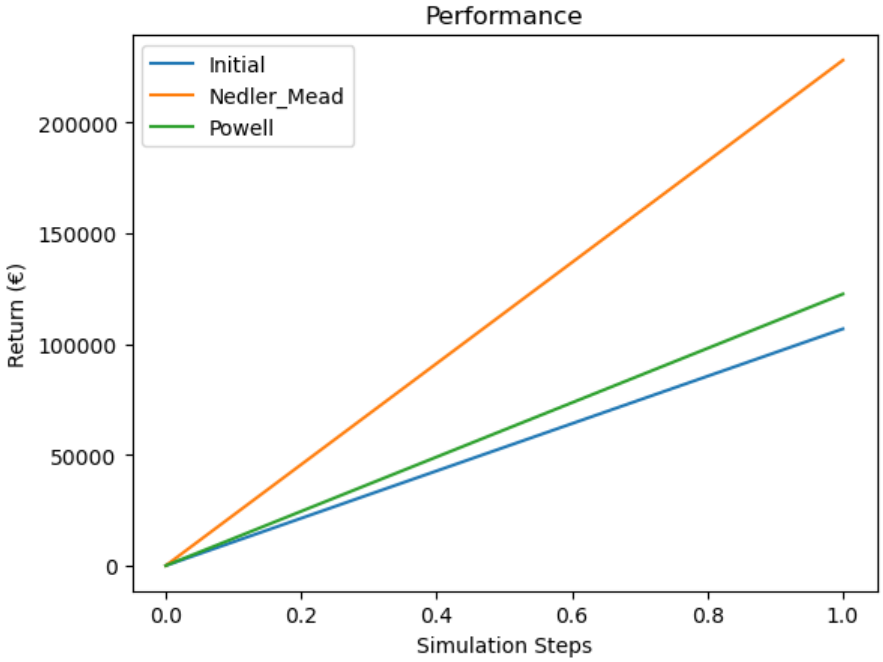
Σε πρώτο στάδιο, για την αρχική εκτίμηση του συστήματος και την καλύτερη προσέγγιση των μεθόδων με τις κατάλληλες παραμέτρους, τόσο η εκπαίδευση όσο και έλεγχος του Πράκτορα πραγματοποιήθηκε για ένα μικρό μέρος των δεδομένων ενώ στην συνέχεια αξιοποιήθηκε όλη η διαθέσιμη πληροφορία. Αρχικά εκπαιδεύτηκε για το τρίμηνο του έτους 2015 των δεδομένων και ο έλεγχος εφαρμόστηκε στον αμέσως επόμενο μήνα της εκπαίδευσης.

Πρωτίστως δημιουργήθηκε μια μέθοδος που λειτουργεί με βάση μια καθορισμένη στρατηγική, χωρίς την χρήση κάποιας τεχνικής Ενισχυτικής Μάθησης. Πρόκειται για έναν βασικό αλγόριθμο, που εκφράζει το επίπεδο αναφοράς (baseline), με βάση το οποίο θα αξιολογηθούν και οι υπόλοιποι Πράκτορας που λειτουργούν σύμφωνα με τις τεχνικές της Ενισχυτικής Μάθησης.

Η αρχή λειτουργίας του για την επιλογή της κάθε ενέργειας βασίζεται σε δυο κατώφλια (thresholds) για την τιμή ανισσοροπίας. Όταν η τιμή της ακριβώς προηγούμενης χρονικής περιόδου ξεπεράσει ένα άνω κατώφλι (upper threshold), ο Πράκτορας αποφασίζει να παραχωρήσει στο δίκτυο την μέγιστη δυνατή ποσότητα ισχύος για το δεδομένο διάστημα dt. Με την ίδια λογική όταν η τιμή ανισορροπίας είναι μικρότερη από ένα ελάχιστο όριο (lower threshold), ο Πράκτορας φορτίζει το σύστημα μπαταρίας με την μέγιστη δυνατή ισχύ, ενώ για οποιαδήποτε ενδιάμεση τιμή, μένει αδρανής. Η συνάρτηση επιλογής αποφάσεων φαίνεται παρακάτω.

$$A_t = \begin{cases} 1 & \lambda^{SI}_{t-1} \leq lower\ threshold \\ -1 & \lambda^{SI}_{t-1} \geq upper\ threshold \\ 0 & otherwise \end{cases}$$

Σε πρώτο στάδιο επιλέχθηκαν αυθαίρετα οι τιμές για το κάθε κατώφλι απόφασης με σκοπό την δημιουργία ενός αρχικού Πράκτορα (Initial). Μια τέτοια μέθοδος, παρά το γεγονός ότι εκμεταλλεύεται τις πολύ υψηλές ή αντίστοιχα τις πολύ χαμηλές τιμές της ανισορροπίας, για να γίνει πιο αποδοτική πρέπει να λειτουργεί για τις βέλτιστες τιμές κατωφλίων. Θεωρώντας ως παραμέτρους το άνω και κάτω κατώφλι εφαρμόστηκαν οι μέθοδοι βελτιστοποίησης Powell και Nelder–Mead για να γίνει προσέγγιση των τιμών που προσφέρουν το μέγιστο συνολικό αποτέλεσμα. Τα αποτελέσματα για τον ορισμό των παραμέτρων φαίνονται στον πίνακα 5.2 και το διάγραμμα 5.1.



Σχήμα 5.1: *Threshold Agents*

Μέθοδος	Lower Threshold	Upper Threshold	Αποτέλεσμα
Initial	-150	150	106771
Powell	-137.2	140	122490
Nelder–Mead	**-195.8**	**89.2**	**227965**

Πίνακας 5.2: *Αποτελέσματα βασικής στρατηγικής*

Όπως φαίνεται, δεδομένου των παραμέτρων της μεθόδου Nelder–Mead, τα τελικά μηνιαία κέρδη για τον διαχειριστή με μια απλή στρατηγική μπορούν να φτάσουν λίγο παραπάνω από 200.000€, ποσό το οποίο είναι κατά πολύ μεγαλύτερο της αρχικής προσέγγισης. Έχοντας ως βάση τα παραπάνω αποτελέσματα, στην συνέχεια έγινε εφαρμογή των μεθόδων Ενισχυτικής Μάθησης.

Το πρώτο βήμα για την εφαρμογή οποιουδήποτε αλγορίθμου Ενισχυτικής Μάθησης που χρησιμοποιεί συναρτήσεις αξίας για την λειτουργία του, είναι να κατασκευαστεί ένας πίνακας που θα εκφράζει την αξία κάθε πιθανής ενέργειας για την κάθε κατάσταση που μπορεί να βρεθεί ο Πράκτορας.

Για να είναι εφικτή η κατασκευή του, αποτελεί προϋπόθεση το Περιβάλλον να επιστρέφει διακριτές τιμές καταστάσεων και ενεργειών από έναν πεπερασμένο χώρο που παίρνουν τις αντίστοιχες τιμές. Έτσι σε πρώτο στάδιο, διακριτοποιήθηκε ο χώρος καταστάσεων και ενεργειών σε n_s και n_α ίσα τμήματα αντίστοιχα.

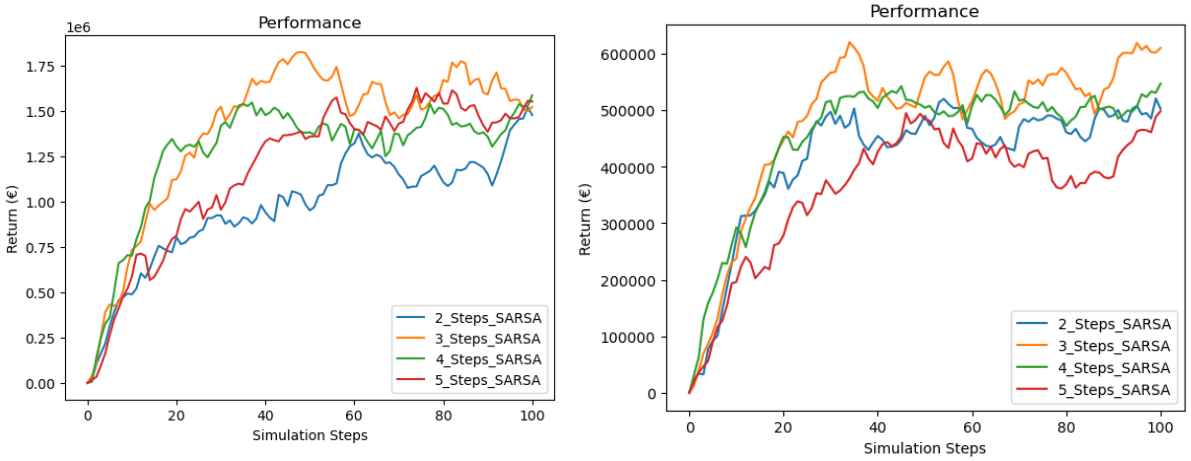
Επίσης σε μια προσπάθεια βελτίωσης της απόδοσης, προστέθηκε μια επιπλέον λειτουργία του Περιβάλλοντος για να μπορέσει ο Πράκτορας να εκπαιδεύεται και να λειτουργεί καλύτερα. Κατασκευάστηκε μια μορφή μνήμης για τον χώρο καταστάσεων έτσι ώστε ο Πράκτορας να μην στηρίζεται μόνο στα δεδομένα της τρέχουσας χρονικής στιγμής. Δημιουργήθηκε μια στοίβα που αποτελείται από πλαίσια καταστάσεων (frames), ο αριθμός των οποίων δηλώνει την ποσότητα των παρελθοντικών καταστάσεων που θα συμπεριληφθούν μαζί με την τρέχουσα.

Παρατηρήθηκε ότι οι μεγάλες τιμές της διακριτοποίησης και του μεγέθους της στοίβας αυξάνουν αντίστοιχα την πολυπλοκότητα του συστήματος και τον χρόνο εκτέλεσης της προσομοίωσης. Έτσι επιλέχθηκε ένας συνδυασμός χαμηλών τιμών που προσφέρει έναν ικανοποιητικό βαθμό απόδοσης. Όλες οι παράμετροι εισαγωγής για το πρώτο μέρος της προσομοίωσης φαίνονται στον πίνακα 5.3.

Παράμετροι		
Πράκτορας	Ρυθμός Μάθησης α Παράγοντας Μείωσης γ *epsilon* Επεισόδια	0.1 0.99 0.1 1000
Περιβάλλον	Αριθμός των frames n_α n_s *dt*	2 5 5 15min

Πίνακας 5.3: *Παράμετροι α΄ μέρους πειράματος*

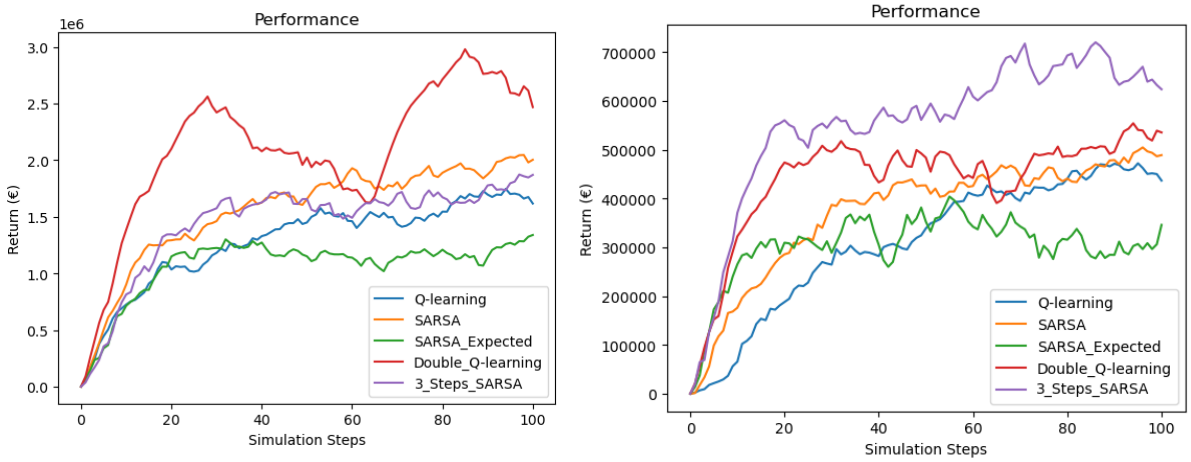
Αρχικά εξετάστηκε η απόκριση του αλγορίθμου n-steps SARSA για διαφορετικές τιμές του n. Πραγματοποιήθηκε προσομοίωση για $n = 1, 2, 3, 4, 5$ και οι ανταμοιβές υπολογίστηκαν μετά από κάθε ολοκλήρωση 10 επεισοδίων τόσο κατά την εκπαίδευση όσο και και κατά τον έλεγχο. Όπως φαίνεται και από τα διαγράμματα 5.2 οι αλγόριθμοι έχουν παρόμοια συμπεριφορά. Όπως είναι αναμενόμενο, το συνολικό κέρδος κατά την εκπαίδευση είναι πολύ μεγαλύτερο σε σχέση αυτό του ελεγχου καθώς πραγματοποιείται για μεγαλύτερο χρονικό διάστημα. Παρατηρείται ότι για $n = 3$ ο Πράκτορας αποδίδει ελαφρώς καλύτερα έναντι των υπολοίπων και γι αυτό τον λόγο, στα επόμενα πειράματα χρησιμοποιήθηκε μόνο ο 3-steps SARSA.



Train *Test*

Σχήμα 5.2: *N-step SARSA Agents*

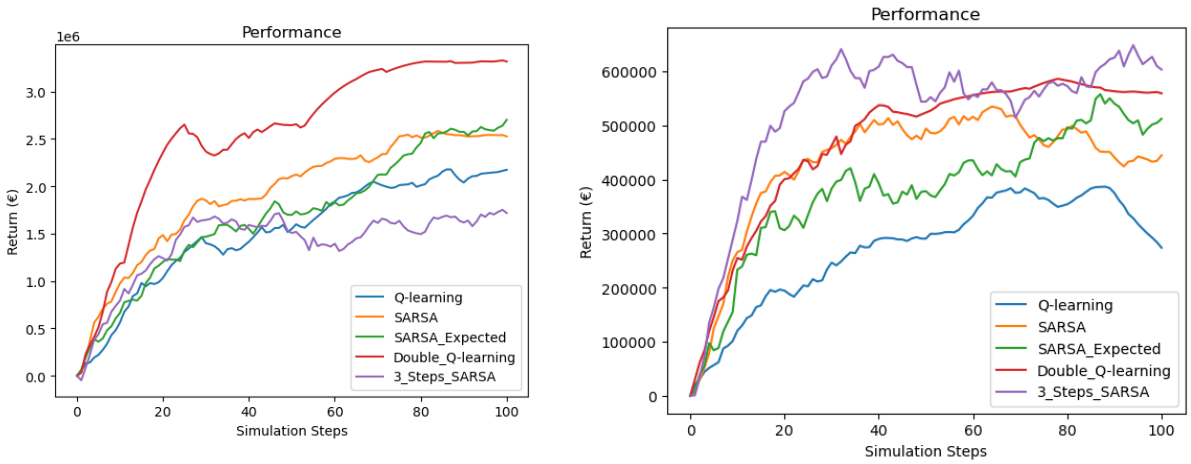
Στη συνέχεια μελετήθηκαν όλοι οι υπόλοιποι αλγόριθμοι και παρουσιάστηκαν στο κοινό διάγραμμα 5.3. Όπως φαίνεται, το κοινό χαρακτηριστικό που έχουν όλοι είναι ότι αποδίδουν πολύ καλύτερα από την βασική στρατηγική αφού το τελικό μηνιαίο κέρδος ξεπερνάει κατά πολύ το ποσό που είχε συγκεντρωθεί για τον ίδιο μήνα. Αξίζει να σημειωθεί ότι ο 3-steps SARSA και ο Double Q-learning εκτός από την ταχύτερη σύγκλιση έναντι των υπολοίπων, επιστρέφουν και τα περισσότερα κέρδη αφού προκύπτουν να είναι τουλάχιστον δύο φορές περισσότερα από την βασική στρατηγική. Από την άλλη πλευρά ο Q-learning με τον SARSA φαίνεται να έχουν παρόμοια συμπεριφορά με τον Double Q-learning, με την διαφορά ότι ο πρώτος συγκλίνει πιο αργά από όλους. Τέλος παρά την γρήγορη του σύγκλιση ο λιγότερο αποδοτικός φαίνεται να είναι ο Expected SARSA.



Train *Test*

Σχήμα 5.3: *Απόδοση αλγορίθμων σε κοινό διάγραμμα*

Στην προσπάθεια για να είναι μικρότερες οι διακυμάνσεις των αποτελεσμάτων, έγινε επανάληψη του ίδιου πειράματος με την διαφορά ότι ο ρυθμός μάθησης φθίνει γραμμικά με την πάροδο των επεισοδίων. Όπως φαίνεται στο 5.4 οι καμπύλες έχουν ελαφρώς ομαλοποιηθεί ενώ ο αλγόριθμος Expected SARSA αποδίδει εμφανώς πολύ καλύτερα από προηγουμένως.

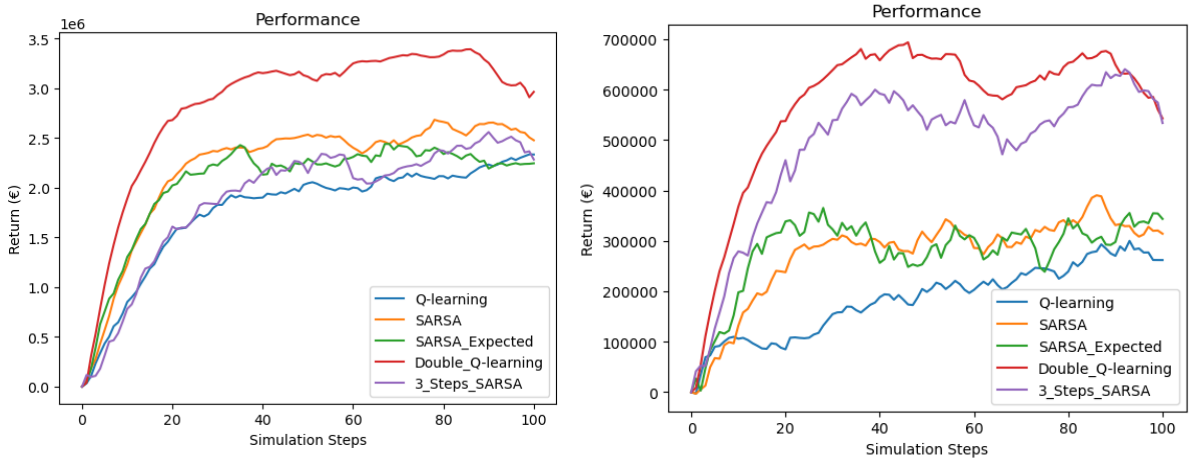


Train *Test*

Σχήμα 5.4: *Απόδοση αλγορίθμων με γραμμική μείωση του ρυθμού μάθησης*

Στην συνέχεια πραγματοποιήθηκε η ίδια προσομοίωση, με την στοίβα αυτή την φορά να αποτελείται από 3 frames.

Για την καλύτερη αξιολόγηση της απόδοσης των αλγορίθμων είναι ανάγκη να υπολογιστούν οι αναπροσαρμοσμένες από τον κίνδυνο χρηματικές απολαβές για την κάθε περίπτωση, μέσω του Sharpe Ratio. Για τον υπολογισμό του, χρησιμοποιήθηκε η baseline στρατηγική ως η



Train *Test*

Σχήμα 5.5: *Απόδοση αλγορίθμων σε κοινό διάγραμμα για περισσότερα frames*

risk-free μέθοδος, σύμφωνα με την σχέση (5.1).

$$SharpeRatio = \frac{R_{method} - R_{baseline}}{\sqrt{var[R_{method} - R_{baseline}]}} \tag{5.1}$$

Από τον πίνακα 5.4 παρατηρείται ότι όλοι αλγόριθμοι αποδίδουν καλύτερα για 3 frames σε

Sharpe Ratios			
Αλγόριθμος	2 frames	2 frames & linear decay	3 frames & linear decay
Q-learning	1.49	1.1	0
Double Q-learning	1.76	**4.1**	3.32
SARSA	1.9	2.38	0.77
SARSA Expected	0.54	1.4	0.46
3-steps SARSA	2.29	2.22	2

Πίνακας 5.4: *Sharpe Ratios για το α΄ μέρος του πειράματος*

συνδυασμό με γραμμική μείωση του ρυθμού μάθησης. Επίσης σε αυτή την περίπτωση φαίνεται ότι ο Double Q-learning έχει εμφανώς την καλύτερη απόδοση έναντι των υπολοίπων.

5.1.2 Β΄ Μέρος πειράματος

Μετά την πρώτη προσέγγιση στο Α΄ μέρους, έγινε επανάληψη της προσομοίωσης, αξιοποιώντας αυτή την φορά όλα τα διαθέσιμα δεδομένα. Πιο συγκεκριμένα έγινε εκπαίδευση για τα έτη 2015-2016 εκτός του τελευταίου μήνα, στον οποίο πραγματοποιήθηκε ο έλεγχος. Οι παράμετροι εισόδου φαίνονται στον πίνακα 5.5.

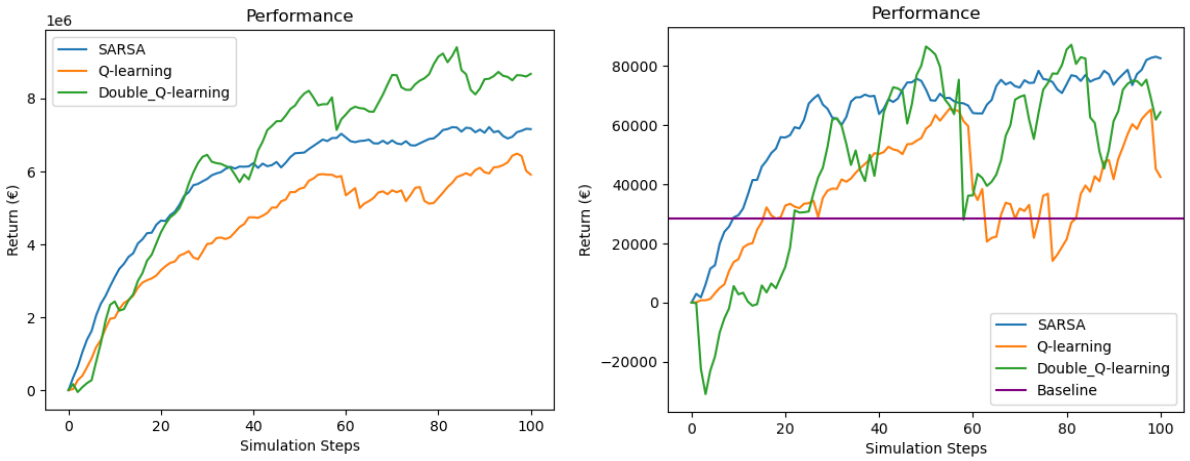
Παράμετροι		
Πράκτορας	Ρυθμός Μάθησης α	0.1
	Παράγοντας Μείωσης γ	0.99
	epsilon	0.1
	Επεισόδια	500
Περιβάλλον	Αριθμός των frames	3
	n_α	3
	n_s	3
	dt	15min

Πίνακας 5.5: *Παράμετροι β΄ μέρους πειράματος*

Για το β΄ μέρος του πειράματος, η εκπαίδευση πραγματοποιήθηκε για τους αλγορίθμους SARSA, Q-learning και Double Q-learning. Αφού έγινε εφαρμογή σε διαφορετικές χρονικές περιόδους από αυτές του α΄ μέρος εκτιμήθηκε και πάλι η απόδοση της baseline στρατηγικής για να είναι ακριβέστερη η εκτίμησή τους. Από τα διαγράμματα 5.6 και τον πίνακα 5.6 φαίνεται και παλι ότι οι αλγόριθμοι Ενισχυτικής Μάθησης ανταποκρίνονται καλύτερα σε βάθος χρόνου από την βασική στρατηγική. Πιο συγκεκριμένα ο SARSA καταφέρνει να πετύχει έως και το διπλάσιο οικονομικό αποτέλεσμα έναντι της baseline και το υψηλότερο Sharpe Ratio. Ωστόσο παρατηρείται ότι ο Q-learning με τον Double Q-learning υστερούν σε απόδοση σε αυτήν την περίπτωση. Αυτό είναι πιθανό να οφείλεται από την παγίδευση των αλγορίθμων σε κάποιο τοπικό ελάχιστο κατά την εκπαίδευση.

Sharpe Ratios	
Q-learning	0.51
Double Q-learning	0.47
SARSA	**2**

Πίνακας 5.6: *Sharpe Ratios για το β΄ μέρος του πειράματος*



Train *Test*

Σχήμα 5.6: *Απόδοση αλγορίθμων για το β΄ μέρος του πειράματος σε κοινό διάγραμμα*

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Σε αυτήν την διπλωματική αναπτύχθηκε ένα σύστημα αποθήκευσης με στόχο να διευκολύνει το έργο στην διαχείριση της μη ισορροπημένης ενέργειας σε μια αγορά. Στόχος ήταν η μέγιστη δυνατή οικονομική απολαβή για τον διαχειριστή του συστήματος αποθήκευσης, που συμμετέχει στην αγορά.

Η Ενισχυτική Μάθηση για τον έλεγχο ενός Συστήματος Αποθήκευσης Ενέργειας παρουσιάζει ιδιαίτερο ενδιαφέρον στον τρόπο με τον οποίο προσεγγίζει την ανάλυσή του. Εκτός του μεγάλου όγκου δεδομένων που μπορεί να διαχειριστεί, είναι σημαντικό να γίνει προσεκτικός σχεδιασμός των διαφορετικών υπολοιήσεων για να μπορέσουν να εκτιμηθούν σωστά όλα τα αποτελέσματα που είναι ικανή να επιφέρει. Αξίζει να σημειωθεί ότι η επιλογή των παραμέτρων αποτελεί πολύ σημαντικό παράγοντα για τον κάθε αλγόριθμο καθώς επηρεάζει άμεσα την απόδοσή του και την στρατηγική που θα ακολουθήσει.

Η στρατηγική του βασικού αλγόριθμου στηρίζεται σε μία απλή μέθοδο που εφαρμόζεται με μεγάλη ευκολία. Αποτελεί μια κοινή διαδικασία που εξασφαλίζει ένα βασικό χρηματικό κέρδος για τον διαχειριστή. Ωστόσο, παρατηρήθηκε ότι οι αλγόριθμοι Ενισχυτικής Μάθησης που εφαρμόστηκαν ξεπερνούν κατά πολύ την απόδοση που θα είχε μια τέτοια απλή προσέγγιση για την διαχείριση του συστήματος. Ενδεικτικά, ο αλγόριθμος SARSA με τις κατάλληλες τροποποιήσεις, καθώς και ο Double Q-learning με αντίστοιχες τεχνικές, αποδίδουν πολύ καλύτερα από κάθε άλλο αλγόριθμο που υλοποιήθηκε και σε ορισμένες περιπτώσεις φάνηκε να καταλήγουν στα διπλάσια χρηματικά αποτελέσματα από την αρχική μέθοδο και να πετυχαίνουν πολύ ικανοποιητικές τιμές του Sharpe Ratio.

6.2 Μελλοντικές Επεκτάσεις

Στα πλαίσια αυτής της διπλωματικής εργασίας η προσέγγιση της διαχείρισης του συστήματος αποθήκευσης φαίνεται να λειτουργεί ικανοποιητικά και τελικά να μπορεί να έχει καλύτερη απόδοση από μια συμβατική στρατηγική. Ωστόσο η ανάπτυξή του μπορεί να επεκταθεί ακόμα περισσότερο.

Αρχικά μπορεί να αναπτυχθεί περαιτέρω το περιβάλλον, υπολογίζοντας περισσότερες χαρακτηριστικά τα οποία θα το κάνουν πιο ρεαλιστικό και αλλά και πιο σύνθετο. Για παράδειγμα μπορεί να λαμβάνει υπόψη του κύκλους φόρτισης της μπαταρίας με σκοπό να εκτιμάται η διάρκεια ζωής του συστήματος κατά την προσομοίωση. Ακόμα θα ήταν χρήσιμο να συνυπολογιστεί η ζήτηση της ενέργειας, καθώς θα επηρεάσει άμεσα την επιλογή των αποφάσεων.

Επίσης η εμφάνιση εποχικών συνιστωσών, ή μεγάλων διακυμάνσεων στα δεδομένα προκαλούν σύγχυση στους αλγορίθμους Ενισχυτικής Μάθησης. Έτσι η προεπεξεργασία των δεδομένων σε συνδυασμό με την επιλογή των κατάλληλων παραμέτρων μπορεί να δημιουργήσει αποδοτικότερους Πράκτορες καθώς δεν θα εμφανίζουν πόλωση στις πολιτικές τους.

Τέλος η χρήση αλγορίθμων Βαθιάς Μηχανικής Μάθησης για την ευκολότερη αξιοποίηση μεγαλύτερου όγκου δεδομένων με σκοπό την ανάπτυξη καινοτόμων και αποτελεσματικών στρατηγικών χωρίς καμία ανθρώπινη καθοδήγηση.

Όλος ο κώδικας υλοποίησης των μεθόδων είναι διαθέσιμος στον σύνδεσμο.

Βιβλιογραφία

[1] Copernicus. *2020 warmest year on record for Europe; globally, 2020 ties with 2016 for warmest year recorded*. https://climate.copernicus.eu/2020-warmest-year-record-europe-globally-2020-ties-2016-warmest-year-recorded, 2021.

[2] NASA. *Global side effects of climate change*. https://climate.nasa.gov/effects/, 2021.

[3] Ioannis Boukas. *Deep Reinforcement Learning for the Control of Energy Storage in Grid-Scale and Microgrid Applications*. Διδακτορική Διατριβή, Université de Liège, Liège, Belgique, 2021.

[4] Directorate Generalfor Energy (European Commission). *Clean energy for all Europeans*. https://op.europa.eu/s/tUCS, 2019.

[5] Dolf Gielen, Ricardo Gorini, Nicholas Wagner, rodrigo leme, Gayathri Prakash, Luca Lorenzoni, Elisa Asmelash, Seán Collins, Luis Janeiro, Rajon Bhuiyan, Rabea Ferroukhi, Michael Renner, Bishal Parajuli, Xavier Casals, Amir Lebdioui, Kelly Rigg, Ulrike Lehr, Eva Alexandri, Unnada Chewpreecha και Pim Vercoulen. *IRENA GRO*. 2020.

[6] Hannah Ritchie. *The price of batteries has declined by 97% in the last three decades*, 2021.

[7] Wesley Cole και A. Will Frazier. *Cost Projections for Utility-Scale Battery Storage*. Τεχνική Αναφορά με αριθμό NREL/TP-6A20-73222, National Renewable Energy Laboratory, 2019.

[8] Niklas Günter και Antonios Marinopoulos. *Energy storage for grid services and applications: Classification, market review, metrics, and methodology for evaluation of deployment cases*. *Journal of Energy Storage*, 8, 2016.

[9] Stuart J. Russell και Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 4η έκδοση, 2020.

[10] Rolandos Alexandros Potamias, Georgios Siolas και Andreas Georgios Stafylopatis. *A transformer-based approach to irony and sarcasm detection*. *Neural Computing and Applications*, 32, 2020.

[11] Rolandos Alexandros Potamias, Georgios Siolas και Andreas Stafylopatis. *A robust deep ensemble classifier for figurative language detection*. τόμος 1000, 2019.

[12] I. Vlahavas, Petros Kefalas, Nick Bassiliades, Fotis Kokkoras και Ilias Sakellariou. *Τεχνητή Νοημοσύνη*. 2020.

[13] Richard S Sutton και Andrew G Barto. *Reinforcement learning: An Introduction (2nd edition 2018)*, τόμος 3. 2018.

[14] Hado Van Hasselt. *Double Q-learning*. 2010.

[15] Ioannis Boukas, Damien Ernst, Thibaut Théate, Adrien Bolland, Alexandre Huynen, Martin Buchwald, Christelle Wynants και Bertrand Cornélusse. *A deep reinforcement learning framework for continuous intraday market bidding. Machine Learning*, 110, 2021.

[16] Gregory Brinkman Paul Denholm, Matthew O'Connell και Jennie Jorgenson. *Overgeneration from Solar Energy in California: A Field Guide to the Duck Chart. National Renewable Energy Laboratory*, 2015.

[17] *Status of Power System Transformation 2018*. https://www.iea.org/reports/status-of-power-system-transformation-2018. 2018, IEA, Παρις.

[18] Niklas Günter και Antonios Marinopoulos. *Energy storage for grid services and applications: Classification, market review, metrics, and methodology for evaluation of deployment cases. Journal of Energy Storage*, 8:226–234, 2016.

[19] Hao Wang και Baosen Zhang. *Energy Storage Arbitrage in Real-Time Markets via Reinforcement Learning. 2018 IEEE Power Energy Society General Meeting (PESGM)*, σελίδες 1–5, 2018.

[20] R. Baxter. *Energy storage*. Ed. Pennwell books, 2006.

[21] Ajay Gambhir Dr Sheridan Few, Oliver Schmidt. *Electrical energy storage for mitigating climate change*. (paper No 20), 2016.

[22] Sandia National Laboratories.U.S. Department of Energy Energy Storage Systems Program. *DOE OE Global Energy Storage Database. Energy*, 2020.

[23] IRENA. *Global energy transformation: A roadmap to 2050*. 2019.

[24] Shafiqur Rehman, Luai M. Al-Hadhrami και Md. Mahbub Alam. *Pumped hydro energy storage system: A technological review. Renewable and Sustainable Energy Reviews*, 44:586–598, 2015.

[25] Goran Strbac Daniel Kirschen. *Fundamentals of Power System Economics*. John Wiley Sons Ltd, University of Manchester Institute of Science Technology (UMIST), UK, 2004.

[26] Λαμπάκης Δημήτριος. *Ελληνικό Χρηματιστήριο Ενέργειας (Target Model)*. Μεταπτυχιακή διπλωματική εργασία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2018.

[27] Κουτσοκούμνης Ν. , Μίλης Μ. *Ανάλυση Ευρωπαϊκών Αγορών Εξισορρόπησης Ηλεκτρικής Ενέργειας*. Πτυχιακή εργασία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2014.

[28] *Ενεργών*. *ΑΔΜΗΕ*, Τεύχος 6, 2018.

[29] Jérémie Bottieau, Louis Hubert, Zacharie De Grève, François Vallée και Jean François Toubeau. *Very-Short-Term Probabilistic Forecasting for a Risk-Aware Participation in the Single Price Imbalance Settlement. IEEE Transactions on Power Systems*, 35(2):1218–1230, 2020.

[30] Simone Totaro, Ioannis Boukas, Anders Jonsson και Bertrand Cornélusse. *Lifelong control of off-grid microgrid with model-based reinforcement learning. Energy*, 232:121035, 2021.

[31] Asmae Berrada και Khalid Loudiyi. *Optimal modeling of energy storage system. International Journal of Modeling and Optimization*, 5(1):71, 2015.

[32] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang και Wojciech Zaremba. *OpenAI Gym*, 2016.