



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Σχεδιασμός Ενός Δίκαιου και Κατανεμημένου Σχήματος Ελέγχου Πολλαπλών Μπαταριών σε Πραγματικό Χρόνο με Χρήση Ενισχυτικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΓΚΟΤΖΙΑ ΓΕΩΡΓΙΟΥ

Επιβλέπουσα: Ελένη Στάη

Επικουρη Καθηγήτρια, Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Σχεδιασμός Ενός Δίκαιου και Κατανεμημένου Σχήματος Ελέγχου Πολλαπλών Μπαταριών σε Πραγματικό Χρόνο με Χρήση Ενισχυτικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ

ΓΚΟΤΖΙΑ ΓΕΩΡΓΙΟΥ

Επιβλέπουσα: Ελένη Στάη

Επίκουρη Καθηγήτρια, Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11η Οκτωβρίου 2024.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....
Ελένη Στάη
Επίκουρη Καθηγήτρια, Ε.Μ.Π.

.....
Συμεών Παπαθασιλείου
Καθηγητής, Ε.Μ.Π.

.....
Βασίλειος Καρυώτης
Αναπλ. Καθηγητής, Ιόνιο Πανεπιστήμιο

Αθήνα, Οκτώβριος 2024



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Copyright ©– Γεώργιος Γκοτζιάς, 2024.

All rights reserved. Με την επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

(Υπογραφή)

.....

Γεώργιος Γκοτζιάς

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η ενσωμάτωση των ανανεώσιμων πηγών ενέργειας πραγματοποιείται με ραγδαίους ρυθμούς τα τελευταία χρόνια. Όμως, η παραγωγή τους χαρακτηρίζεται από αβεβαιότητες και στοχαστικότητα, δυσχεραίνοντας τις προβλέψεις, με αποτέλεσμα την ανάγκη για αυξημένα αποθέματα, τα οποία ωστόσο αυξάνουν τα λειτουργικά κόστη. Ως εναλλακτικός τρόπος αποθήκευσης ενέργειας προτείνεται η χρήση μπαταριών. Οι μπαταρίες μπορούν να αποθηκεύουν ή να προσφέρουν ενέργεια, ώστε να εξισορροπηθεί η προσφορά με τη ζήτηση. Έτσι, δημιουργείται η ανάγκη για βέλτιστο έλεγχο τους.

Λαμβάνοντας υπόψιν ότι κάθε μπαταρία έχει περιορισμούς για την ασφαλή λειτουργία της, καθώς και το συνολικό δίκτυο μπορεί να έχει στόχους σχετικά με την αποθηκευμένη ενέργεια συγκεκριμένη χρονική στιγμή, το πρόβλημα του ελέγχου των πολλαπλών μπαταριών ανάγεται σε ένα πρόβλημα βελτιστοποίησης. Παραδοσιακά, τέτοια προβλήματα επιλύονται βέλτιστα με τη μέθοδο Model Predictive Control. Τα τελευταία χρόνια, η αυξανόμενη χρήση της μηχανικής μάθησης έχει οδηγήσει σε μεθόδους της να βρίσκουν εφαρμογές και στον βέλτιστο έλεγχο έξυπνων δικτύων. Συγκεκριμένα, η ενισχυτική μάθηση είναι το πεδίο της μηχανικής μάθησης που επιχειρεί να ανακαλύψει τη βέλτιστη πολιτική βάσει της αλληλεπίδρασης με το περιβάλλον. Το γεγονός ότι λειτουργεί υπό στοχαστικότητα την καθιστά κατάλληλη για τέτοιες εφαρμογές, ενώ παρέχονται μέθοδοι που επιτρέπουν την κατανεμημένη απόφαση κάθε μπαταρίας, έναντι κεντρικών αποφάσεων αυξάνοντας την ιδιωτικότητα.

Σε κατανεμημένα περιβάλλοντα εγείρονται ζητήματα σχετικά με τη δίκαιη χρήση των μπαταριών. Ενδεικτικά, μπορεί μία μπαταρία να μη χρησιμοποιείται, με αποτέλεσμα οι υπόλοιπες να φορτίζουν ή ξεφορτίζουν περισσότερο από όσο θα χρειαζόταν. Στην παρούσα εργασία, προτείνονται μέθοδοι, σύμφωνα με τις οποίες, ο βέλτιστος έλεγχος των μπαταριών με χρήση του αλγορίθμου ενισχυτικής μάθησης πολλαπλών πρακτόρων MADDPG καταφέρνει να εκμάθει για κάθε μπαταρία μία πολιτική, η οποία ελαχιστοποιεί το κόστος του δικτύου, ενώ ταυτόχρονα επιτυγχάνει τη δίκαιη χρήση μεταξύ των μπαταριών. Συγκεκριμένα, εξετάζονται δύο τρόποι με τους οποίους επιτυγχάνεται η δικαιοσύνη μεταξύ των μπαταριών, ενώ εξετάζεται και η πληροφορία που πρέπει να κοινοποιείται από το δίκτυο με σκοπό τον βέλτιστο έλεγχο.

Λέξεις Κλειδιά

Έξυπνα δίκτυα ενέργειας, Συστήματα μπαταριών αποθήκευσης ενέργειας, Πολυ-πρακτορική ενισχυτική μάθηση, Κυρτή Βελτιστοποίηση, Κριτήρια Δικαιοσύνης

Abstract

The use of Renewable Energy Sources (RES) has been rapidly increasing in recent years. However, the production of RES is uncertain and stochastic, making it difficult to predict and requiring additional reserves, which increases operating costs. An alternative solution for energy storage is the use of batteries. Batteries can store or provide energy to balance the supply and the demand. Thus, control of the batteries must be optimal.

Taking into consideration the safety constraints that each battery has to satisfy and the additional targets that the grid has in total regarding the total reserves in specific timesteps, the optimal control problem is reduced to an optimization problem. The most common method of solving this problem is Model Predictive Control. Lately, the use of machine learning methods is expanding rapidly and consequently, some methods find applications in optimal control of smart grids. Specifically, reinforcement learning is the field of machine learning that discovers an optimal policy for a problem by interacting with the environment. The fact that it performs even under stochasticity makes it suitable for such applications, additionally, it provides methods for the distributed control of each battery instead of the typical centralized decisions, thereby increasing privacy.

In distributed approaches, concerns are raised regarding the fair use of the batteries. For instance, a battery could remain unused, meanwhile another battery has to charge or discharge more to compensate. In this work, we provide methods based on the multi-agent reinforcement learning algorithm, MADDPG, which leads to the optimal control of batteries, while ensuring equitable usage among the batteries. In particular, we compare two methods to provide fairness and examine two possible state representations regarding the information that should be shared by the grid to optimize battery control.

Keywords

Smart grids, Battery Energy Storage Systems, Multi-agent Reinforcement Learning, Convex Optimization, Fairness criteria

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια αυτής της διπλωματικής, κυρία Ελένη Στάη, για τις συμβουλές, την καθοδήγηση, τον χρόνο που αφιέρωσε και το συνεχές ενδιαφέρον της για την πορεία της εργασίας. Η συμβολή της για την ολοκλήρωση της εργασίας ήταν καθοριστική.

Ακόμη, οφείλω να ευχαριστήσω τον κύριο Συμεών Παπαθασιλείου, που δέχτηκε να πραγματοποιήσω τη διπλωματική μου εργασία στο εργαστήριό του, καθώς και για τη συμβολή του στην επιλογή θέματος.

Τέλος, ευχαριστώ την οικογένειά μου και τους φίλους που με στήριξαν καθ' όλη τη διάρκεια των σπουδών μου.

Αθήνα, Οκτώβριος 2024

Γεώργιος Γκοιζιάς

Περιεχόμενα

Περίληψη	1
Abstract	3
Ευχαριστίες	5
1 Εισαγωγή	15
1.1 Αντικείμενο της διπλωματικής	16
1.2 Σχετική Βιβλιογραφία	17
1.3 Οργάνωση του τόμου	18
2 Βελτιστοποίηση Σε Έξυπνα Δίκτυα Ενέργειας- Κριτήρια Δικαιοσύνης	19
2.1 Κυριτή Βελτιστοποίηση	19
2.2 Συμβολισμοί	20
2.3 Υπολογισμός Τροχιών Κατάστασης Ενέργειας Μπαταριών και Πλάνου Διανομής	21
2.4 Δίκαιος Έλεγχος Μπαταριών	22
2.4.1 Συνάρτηση κόστους μπαταρίας	23
2.4.2 Συνάρτηση κόστους δικτύου- Υπολογισμός βαρών	25
2.5 Κριτήρια Δικαιοσύνης	27
2.5.1 Δικαιοσύνη βάσει των περιορισμών ισχύος	27
2.5.2 Δικαιοσύνη βάσει της τρέχουσας κατάστασης της ενέργειας	28
3 Μηχανική Μάθηση	31
3.1 Μηχανική Μάθηση	31
3.1.1 Ορισμός	31
3.1.2 Κατηγορίες Μηχανικής Μάθησης	32
3.2 Νευρωνικά Δίκτυα	33
3.2.1 Δομή Νευρωνικών Δικτύων	33
3.2.2 Αλγόριθμος Perceptron	34
3.2.3 Αρχιτεκτονικές Νευρωνικών Δικτύων	34
3.3 Βαθιά Μάθηση	35
3.3.1 Δίκτυα πρόσθιας τροφοδότησης	35
3.3.2 Συνάρτηση ενεργοποίησης	36
3.3.3 Εκπαίδευση MLP	36
3.4 Μαρκοβιανές Διαδικασίες Απόφασης	37
3.4.1 Θεωρία Μαρκοβιανών Διαδικασιών Απόφασης	37

3.4.2	Μαρκοβιανά Παίγνια	38
3.5	Ενισχυτική Μάθηση	39
3.5.1	Βασικοί Ορισμοί	39
3.5.2	Κατηγορίες Ενισχυτικής Μάθησης	40
3.5.3	Εκμάθηση Συναρτήσεων	41
3.5.4	Policy Gradient	43
3.6	Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων	47
3.6.1	Κατηγορίες MARL	47
3.6.2	Multi-Agent Deep Deterministic Policy Gradient (MADDPG)	48
4	Μοντελοποίηση προβλήματος	51
4.1	Μαθηματική Περιγραφή Προβλήματος	51
4.2	Σχεδιασμός Δίκαιων Τροχιών	55
4.2.1	Δίκαιες τροχιές βάσει των περιορισμών ισχύος	55
4.2.2	Δίκαιες τροχιές βάσει της τρέχουσας Κατάστασης της Ενέργειας	57
4.3	Αναπαράσταση της Κατάστασης	60
4.4	Συνάρτηση Ανταμοιβής βάσει Τροχιών	61
4.5	Συνάρτηση Ανταμοιβής χωρίς Τροχιές	68
4.6	Εφαρμογή του αλγορίθμου MADDPG	69
5	Υλοποίηση	75
5.1	Λεπτομέρειες Υλοποίησης	75
5.2	Περιπτώσεις Χρήσης	76
5.2.1	Σενάρια Εκπαίδευσης	76
5.2.2	Περίπτωση 1: 2 Μπαταρίες και Βέλτιστο Πλάνο Διανομής	77
5.2.3	Περίπτωση 2: Αυτόνομη λειτουργία δικτύου 2 μπαταριών	77
5.2.4	Περίπτωση 3: Αυτόνομη λειτουργία δικτύου 4 μπαταριών	78
5.3	Μέθοδοι Προς Αξιολόγηση	78
5.4	Επιλογή Παραμέτρων	81
5.5	Πειράματα	82
5.5.1	Περίπτωση 1	83
5.5.2	Περίπτωση 2	92
5.5.3	Περίπτωση 3	97
6	Επίλογος	105
6.1	Συμπεράσματα	105
6.2	Αντικείμενα Μελλοντικής Έρευνας	106
	Παραρτήματα	107
Α΄	Υπολογισμός Τροχιών και Δικαιοσύνη	109
Α.1	Αγνοώντας το battery cycling cost	109
Α.2	Θεωρώντας το battery cycling cost	110
Α.3	Απόδειξη ικανοποίησης των συνδυαστικών περιορισμών από τον αλγόριθμο 4.1	111

Β' Υπολογισμός Βαρών	113
Β'.1 Περιγραφή Άνω Φράγματος	113
Β'.2 Απόδειξη γενικότερης πρότασης	115
Β'.3 Μέθοδος 2.4 και Κριτήριο Δικαιοσύνης 2.5.1	117
Γ' Παράμετροι	119
Βιβλιογραφία	121
Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια	125

Κατάλογος Σχημάτων

3.1	Τεχνητό Νευρωνικό Δίκτυο	34
3.2	Αρχιτεκτονική Δικτύου Perceptron Πολλαπλών Επιπέδων	35
3.3	Αρχιτεκτονική Actor-Critic	44
5.1	Πιθανά Σενάρια Εκπαίδευσης	76
5.2	Πλάνο Διανομής	78
5.3	Μέσο Συνολικό Κόστος κατά την εκπαίδευση (1η περίπτωση)	84
5.4	Μέση Συνολική Ανταμοιβή κατά την εκπαίδευση (1η περίπτωση)	84
5.5	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 1)	85
5.6	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 1)	85
5.7	Φορτίο και απόκλιση από το πλάνο διανομής στο PCC (Περίπτωση 1)	86
5.8	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 2)	86
5.9	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 2)	86
5.10	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 3)	87
5.11	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 3)	87
5.12	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 4)	88
5.13	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 4)	88
5.14	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 5)	88
5.15	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 5)	88
5.16	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 6)	89
5.17	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 6)	89
5.18	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 7)	89
5.19	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 7)	89
5.20	Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 8)	90
5.21	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 8)	90
5.22	Μέσο Συνολικό Κόστος κατά την εκπαίδευση (2η περίπτωση)	92
5.23	Μέση Συνολική Ανταμοιβή κατά την εκπαίδευση (2η περίπτωση)	92
5.24	Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 1)	93
5.25	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 1)	93
5.26	Φορτίο και απόκλιση από το πλάνο διανομής στο PCC (Περίπτωση 2)	94
5.27	Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 3)	94
5.28	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 3)	94
5.29	Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 4)	95
5.30	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 4)	95
5.31	Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 5)	95

5.32	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 5)	95
5.33	Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 8)	96
5.34	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 8)	96
5.35	Μέσο Συνολικό Κόστος κατά την εκπαίδευση (3η περίπτωση)	98
5.36	Μέση Συνολική Ανταμοιβή κατά την εκπαίδευση (3η περίπτωση)	98
5.37	Φορτίο και απόκλιση από το πλάνο διανομής στο PCC (Περίπτωση 3)	99
5.38	Ενεργός ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 2)	100
5.39	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 2)	100
5.40	Ενεργός ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 3)	100
5.41	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 3)	100
5.42	Ενεργός ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 6)	101
5.43	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 6)	101
5.44	Ενεργός ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 7)	102
5.45	Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 7)	102
B'.1	Η συνάρτηση $g(w_m, d)$: Πρώτη περίπτωση	114
B'.2	Η συνάρτηση $g(w_m, d)$: Δεύτερη περίπτωση	114
B'.3	Η συνάρτηση $g(w_m, d)$: Τρίτη περίπτωση	115

Κατάλογος Πινάκων

5.1	Αποτελέσματα κόστους μεθόδων αναφοράς για την περίπτωση 1	85
5.2	Αποτελέσματα μεθόδων MADDPG για το κόστος (Περίπτωση 1)	90
5.3	Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.1 (Περίπτωση 1)	91
5.4	Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.2 (Περίπτωση 1)	91
5.5	Αποτελέσματα κόστους μεθόδων αναφοράς για την περίπτωση 2	92
5.6	Αποτελέσματα μεθόδων MADDPG για το κόστος (Περίπτωση 2)	97
5.7	Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.1 (Περίπτωση 2)	97
5.8	Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.2(Περίπτωση 2)	97
5.9	Αποτελέσματα κόστους μεθόδων αναφοράς για την περίπτωση 3	98
5.10	Αποτελέσματα μεθόδων MADDPG για το κόστος (Περίπτωση 3)	102
5.11	Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.1 (Περίπτωση 3)	102
5.12	Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.2 (Περίπτωση 3)	103
Γ.1	Χαρακτηριστικά μπαταριών (σε ανά μονάδα τιμές)	119
Γ.2	Παράμετροι προβλημάτων βελτιστοποίησης (σε ανά μονάδα τιμές)	119
Γ.3	Τιμές Υπερπαραμέτρων	119

Εισαγωγή

Οι ανανεώσιμες πηγές ενέργειας (ΑΠΕ), όπως τα φωτοβολταϊκά και οι ανεμογεννήτριες, χρησιμοποιούνται όλο και περισσότερο στα σύγχρονα συστήματα ηλεκτρικής ενέργειας. Αυτό παρατηρείται, καθώς οι ΑΠΕ αποτελούν βιώσιμες πηγές ενέργειας. Αντίθετα, παραδοσιακοί μέθοδοι παραγωγής ηλεκτρικής ενέργειας, όπως η καύση λιγνίτη, παράγουν σημαντικές ποσότητες διοξειδίου του άνθρακα, αποτελώντας μία κύρια αιτία για την κλιματική αλλαγή.

Παράλληλα, η ραγδαία ανάπτυξη της τεχνολογίας βρίσκει εφαρμογή και στα δίκτυα ενέργειας. Πλέον, τα συστήματα αυτά χαρακτηρίζονται ως έξυπνα δίκτυα ενέργειας (smart grids), καθώς χρησιμοποιούνται αισθητήρες και μικροελεγκτές, ενώ είναι δυνατή και η επικοινωνία μεταξύ απομακρυσμένων μερών του δικτύου. Η ενσωμάτωση αυτών των τεχνολογιών δημιουργεί νέες δυνατότητες σχετικά με το βέλτιστο έλεγχο των δικτύων ενέργειας.

Στην παρούσα εργασία το πρόβλημα του βέλτιστου ελέγχου εστιάζεται σε ένα μικρό υποσύνολο του συνολικού συστήματος, το οποίο αποτελείται από ΑΠΕ, συμβατικές μονάδες παραγωγής, καθώς και συστήματα αποθήκευσης ενέργειας, τα οποία συνδέονται στο συνολικό σύστημα μέσω του σημείου κοινής σύζευξης (Point of Common Coupling - PCC)[1]. Ο σκοπός του ελέγχου είναι πρακτικά να εξισορροπήσει την προσφορά με τη ζήτηση.

Επειδή η παραγωγή των ΑΠΕ χαρακτηρίζονται από αβεβαιότητα και στοχαστικότητα είναι αναγκαία η χρήση συστημάτων αποθήκευσης ενέργειας, όπως οι μπαταρίες. Οι μπαταρίες είναι πλήρως ελέγξιμες, οπότε μπορούν είτε να αποθηκεύουν ενέργεια για μελλοντική χρήση, όταν η προσφορά ξεπερνά τη ζήτηση, είτε να προσφέρουν ενέργεια, σε αντίθετη περίπτωση.

Ο σύνηθης τρόπος ελέγχου πραγματοποιείται σε τρία επίπεδα [2]:

1. Το πρώτο επίπεδο αφορά τον υπολογισμό ενός σχεδίου διανομής (dispatch plan) για σχετικά μεγάλο χρονικό διάστημα (π.χ. μία ημέρα).
2. Το δεύτερο επίπεδο αφορά έναν μέσο χρονικό ορίζοντα (όπως 15 λεπτά), με σκοπό να διορθώσει τις πιθανώς λανθασμένες προβλέψεις του πρώτου επιπέδου.
3. Το τρίτο επίπεδο αφορά μικρά χρονικά διαστήματα (π.χ. 10 δευτερόλεπτα) και έχει ως σκοπό τον έλεγχο σε πραγματικό χρόνο, για τα βέλτιστα δυνατά αποτελέσματα.

Έτσι, το πρόβλημα του βέλτιστου ελέγχου σε πραγματικό χρόνο, στο οποίο και επικεντρώνεται αυτή η εργασία, εκφράζεται ως το πρόβλημα ελαχιστοποίησης του σφάλματος της παρατηρούμενης ισχύος στο PCC σε σχέση με την ισχύ που προβλέπεται από το πλάνο.

Παραδοσιακά, για τον έλεγχο χρησιμοποιούνταν κεντριοκοιμήμενες μέθοδοι, όμως παρατηρούνται προβλήματα κλιμακωσιμότητας, λόγω των αυξημένων υπολογιστικών πόρων και της επικοινωνίας που απαιτούν, ενώ παρατηρείται και μοναδικό σημείο αποτυχίας. Αντίθετα, η σύγχρονη τάση αφορά των σχεδιασμό κατανεμημένων σχημάτων ελέγχου, τα οποία προσφέρουν κλιμακωσιμότητα και αξιοπιστία, παρ' ότι εξακολουθούν να απαιτούν να υπάρχει επικοινωνία μεταξύ κάποιων μερών [1, 3].

1.1 Αντικείμενο της διπλωματικής

Σκοπός της παρούσας εργασίας είναι ανάπτυξη σχημάτων που αφορά τον βέλτιστο έλεγχο μπαταριών σε ένα έξυπνο δίκτυο ενέργειας λαμβάνοντας υπόψιν τη δίκαιη χρήση μεταξύ των μπαταριών. Ο έλεγχος αφορά το κατώτερο επίπεδο, όπως περιγράφηκε προηγουμένως, δηλαδή το επίπεδο που αφορά τις αποφάσεις σε πραγματικό χρόνο. Στόχος του ελέγχου είναι να παρατηρείται στο PCC η ισχύς του πλάνου διανομής, ενώ πληρούνται οι περιορισμοί ασφάλειας κάθε μπαταρίας, καθώς και περαιτέρω στόχοι του δικτύου, όπως η αποθηκευμένη ενέργεια σε ένα υποσύνολο των μπαταριών να βρίσκεται μεταξύ συγκεκριμένων ορίων σε προκαθορισμένες χρονικές στιγμές.

Συγκεκριμένα, παρουσιάζονται μέθοδοι βασισμένοι στον αλγόριθμο ενισχυτικής μάθησης πολλαπλών πρακτόρων MADDPG, με τις οποίες επιτυγχάνεται ο βέλτιστος έλεγχος των μπαταριών για να ακολουθείται το σχέδιο διανομής, ενώ ταυτόχρονα επιτυγχάνεται η δίκαιη χρησιμοποίηση των μπαταριών, σύμφωνα με κριτήρια που επιλέγονται.

Οι προσεγγίσεις που περιγράφονται στην παρούσα εργασία αφορούν κατανεμημένα σχήματα, για τα οποία είναι αναγκαία η επικοινωνία μεταξύ μπαταρίας και PCC, ενώ δεν χρειάζεται επικοινωνία μεταξύ των μπαταριών, καθώς θα περιορίζονταν η κλιμακωσιμότητα. Αντίθετα, η επικοινωνία με το PCC κρίνεται απαραίτητη, ώστε να λαμβάνεται πληροφορία για την παραγωγή των υπόλοιπων πηγών ενέργειας, με στόχο την ελαχιστοποίηση του κόστους.

Σχετικά, με την επίτευξη της δικαιοσύνης εξετάζονται δύο τρόποι:

1. Ο πρώτος αφορά τη χρήση τροχιών, όπου η δίκαιη χρήση εξασφαλίζεται κατά τον σχεδιασμό τους. Ο λόγος είναι ότι είναι ευκολότερη η εύρεση δίκαιων τροχιών, καθώς προκύπτουν από την ασύγχρονη επίλυση προβλημάτων βελτιστοποίησης.
2. Ο δεύτερος αφορά την εισαγωγή του κριτηρίου δικαιοσύνης απευθείας στην ανταμοιβή που χρησιμοποιείται στην ενισχυτική μάθηση.

Οι μέθοδοι αυτοί εξετάζονται σε διαφορετικά σενάρια χρήσης, καθώς και για διαφορετικό αριθμό μπαταριών, ώστε να γίνουν εμφανή τα πλεονεκτήματα και τα μειονεκτήματα κάθε επιλογής έναντι των υπολοίπων. Τα κριτήρια της σύγκρισης αφορούν την κάλυψη του ισοζυγίου ισχύος στο PCC, την ικανοποίηση των περιορισμών ασφαλείας κάθε μπαταρίας, την ικανοποίηση των περιορισμών σχετικά με τους στόχους του δικτύου, καθώς και μετρικές σχετικές με τη δικαιοσύνη. Τέλος, παραθέτονται συγκριτικά και τα αποτελέσματα που προκύπτουν για τις ίδιες περιπτώσεις από τον αλγόριθμο MADDPG αν δεν χρησιμοποιηθεί κανένα κριτήριο δικαιοσύνης.

1.2 Σχετική Βιβλιογραφία

Το [4] πραγματεύεται το πρόβλημα της εξισορρόπησης του ισοζυγίου ισχύος, όπου η επίλυση πραγματοποιείται μέσω μιας κεντρικοποιημένης προσέγγισης βασισμένη στον αλγόριθμο ενισχυτικής μάθησης Deep Deterministic Policy Gradient (DDPG). Τέτοιες κεντρικοποιημένες προσεγγίσεις παρουσιάζουν προβλήματα σχετικά με την κλιμακωσιμότητας και απαιτούν αυξημένη ανταλλαγή πληροφορίας σε σχέση με τις κατανεμημένες μεθόδους, παρότι παρέχουν καλύτερα αποτελέσματα, αφού υπάρχει πλήρης εικόνα του δικτύου. Η μέθοδος που παρουσιάζεται μπορεί να λάβει υπόψιν το κόστος της ενέργειας, όμως δεν παρέχεται τρόπος, ώστε να εξισορροπείται η χρήση των μπαταριών.

Στην εργασία [5] παρέχεται μια πρώτη μέθοδος επίλυσης του προβλήματος με χρήση ενισχυτικής μάθησης πολλαπλών πρακτόρων. Στην προσέγγιση αυτή, κάθε πράκτορας μαθαίνει ανεξάρτητα από τους υπολοίπους. Χρησιμοποιείται η μέθοδος Lagrangian Decomposition, ώστε το πρόβλημα να αποσυντεθεί σε επιμέρους προβλήματα, τα οποία εμπλέκουν μόνο τον συντελεστή Lagrange και πληροφορία που υπάρχει τοπικά σε κάθε μπαταρία. Η εκπαίδευση των πρακτόρων γίνεται μέσω δεδομένων που προκύπτουν από επίλυση προβλημάτων βελτιστοποίησης βάσει κάποιων σεναρίων και όχι από αλληλεπίδραση με το περιβάλλον. Η συνάρτηση ανταμοιβής, η οποία χρησιμοποιείται για τη λήψη αποφάσεων, προσεγγίζεται βάσει της μεθόδου Fitted Q-iteration.

Στο [6] παρουσιάζεται μια μέθοδος ενισχυτικής μάθησης πολλαπλών πρακτόρων, στην οποία η εκπαίδευση γίνεται κεντρικοποιημένα, ώστε η πολιτική κάθε πράκτορα να λαμβάνει υπόψιν την πολιτική των υπολοίπων, χωρίς όμως να απαιτεί τη γνώση της πολιτικής κατά την εκτέλεση. Αυτή η μέθοδος επιτρέπει την κατανεμημένη λήψη αποφάσεων για κάθε πράκτορα, ενώ σε αντίθεση με προηγούμενες μεθόδους υπάρχει περισσότερη πληροφορία κατά την εκπαίδευση, οπότε είναι δυνατόν να βρεθούν καλύτερες λύσεις.

Η προηγούμενη μέθοδος εφαρμόζεται στο ζητούμενο πρόβλημα στο [7]. Η διαφορά με την παρούσα εργασία έγκειται στη δίκαιη χρήση μπαταριών, για την οποία δεν παρέχεται καμία εγγύηση στην προηγούμενη εργασία. Ακόμη, υπάρχει διαφοροποίηση σχετικά με την αναπαράσταση της κατάστασης του περιβάλλοντος που χρησιμοποιείται. Αυτό επιλέγεται, ώστε να αποφεύγονται τα προβλήματα υπολογιστικής πολυπλοκότητας που παρουσιάζονται με χρήση της τεχνικής Lagrangian Decomposition, αλλά και τα προβλήματα της εναλλακτικής που παρουσιάζεται, η οποία δεν λαμβάνει υπόψιν το τρέχον φορτίο που παράγουν οι ΑΠΕ, όπως αναμένεται στο επίπεδο ελέγχου πραγματικού χρόνου.

Εφαρμογές της ενισχυτικής μάθησης πολλαπλών πρακτόρων σε περιβάλλοντα συστημάτων ενέργειας υπό αβεβαιότητα παρουσιάζονται και στην εργασία [8]. Στο έργο αυτό, χρησιμοποιείται ανεξάρτητη εκπαίδευση των πρακτόρων, όπου κάθε πράκτορας χρησιμοποιεί μόνο πληροφορία που είναι διαθέσιμη τοπικά, με αποτέλεσμα να μην χρειάζεται επικοινωνία μεταξύ των επιμέρους πρακτόρων. Η εργασία αυτή προτείνει τη χρήση πινάκων σταθερού μεγέθους για την αναπαράσταση της συνάρτησης αξίας δράσης (action-value function) της ενισχυτικής μάθησης, με σκοπό τη βελτίωση της κλιμακωσιμότητας. Επίσης, προτείνεται η χρήση της εκπαίδευσης μέσω δεδομένων που προκύπτουν από την επίλυση προβλημάτων βελτιστοποίησης offline, με σκοπό την βελτίωση της εξερεύνησης κατά την αλληλεπίδραση με το περιβάλλον.

Σχετικά με τον αλγόριθμο MADDPG που χρησιμοποιείται και στην παρούσα εργασία, η εφαρμογή του σε smart grids εξετάζεται και στο [9]. Στην περίπτωση αυτή, χρησιμοποιείται για τον έλεγχο της συχνότητας φορτίου. Η προσέγγιση αυτή επιτυγχάνει έλεγχο χρησιμοποιώντας πληροφορία που είναι διαθέσιμη τοπικά στον κάθε πράκτορα. Η εκδοχή αυτή του αλγορίθμου MADDPG κάνει χρήση νευρωνικών δικτύων Long Short-Term Memory (LSTM), ώστε να χρησιμοποιείται για τις αποφάσεις όλη η πληροφορία που έχει παρατηρηθεί ως εκείνο το χρονικό σημείο.

Σχετικά με τη δίκαιη χρήση των μπαταριών, στο [2] παρουσιάζεται ένα κριτήριο δικαιοσύνης σχετικά με την εξισορρόπηση της του κοινού στόχου για το δίκτυο και το στόχο που έχει η κάθε μπαταρία, για να ακολουθεί μία τροχιά. Η εύρεση της τροχιάς, την οποία οφείλει να ακολουθεί η μπαταρία, μπορεί να πραγματοποιηθεί με τη μέθοδο που παρουσιάζεται στο [10]. Το γεγονός ότι η εύρεση γίνεται μέσω επίλυσης προβλήματος βελτιστοποίησης offline καθιστά ευκολότερη την εισαγωγή ενός κριτηρίου δικαιοσύνης στον υπολογισμό των τροχιών.

Στο έργο [11] παρουσιάζεται μία μέθοδος σχετικά με την εξισορρόπηση της φόρτισης των μπαταριών, ώστε να μην φορτίσει ή ξεφορτίσει τελείως κάποια μπαταρία σε σχέση με τις υπόλοιπες. Η μέθοδος που παρουσιάζεται αφορά την επίτευξη του ελέγχου μέσω ενός καταναμημένου αλγορίθμου, κατά τον οποίο η ανταλλαγή πληροφορίας πραγματοποιείται μόνο όταν είναι αναγκαία. Η εξισορρόπηση της κατάστασης της ενέργειας που επιτυγχάνει αυτή η μέθοδος μπορεί να χρησιμοποιηθεί ως ένα κριτήριο δικαιοσύνης.

1.3 Οργάνωση του τόμου

Η παρούσα εργασία ακολουθεί την παρακάτω δομή:

- Στο [Κεφάλαιο 2](#) πραγματεύονται τα κριτήρια δικαιοσύνης, ενώ εξηγείται η θεωρία βελτιστοποίησης και η εφαρμογή της στο σχεδιασμό τροχιών σε έξυπνα δίκτυα ενέργειας, με σκοπό την ενσωμάτωση των κριτηρίων δικαιοσύνης στο πρόβλημα ελέγχου.
- Στο [Κεφάλαιο 3](#) παρουσιάζεται το θεωρητικό υπόβαθρο από την περιοχή της μηχανικής μάθησης.
- Στο [Κεφάλαιο 4](#) το πρόβλημα του βέλτιστου ελέγχου σε πραγματικό χρόνο μοντελοποιείται ως πρόβλημα ενισχυτικής μάθησης.
- Στο [Κεφάλαιο 5](#) περιγράφεται η υλοποίηση μέσω της οποίας πραγματοποιήθηκαν οι προσομοιώσεις, ενώ στη συνέχεια παρουσιάζονται τα σενάρια που εξετάστηκαν, καθώς και τα σχετικά αποτελέσματα.
- Στο [Κεφάλαιο 6](#) αναλύονται τα αποτελέσματα των περιπτώσεων χρήσης που εξετάστηκαν, παρουσιάζονται τα συμπεράσματα που προκύπτουν και δίνονται θέματα προς μελλοντική έρευνα.

Κεφάλαιο 2

Βελτιστοποίηση Σε Έξυπνα Δίκτυα Ενέργειας-Κριτήρια Δικαιοσύνης

Στο κεφάλαιο αυτό αρχικά γίνεται εισαγωγή στη μαθηματική βελτιστοποίηση. Στη συνέχεια, περιγράφεται ο τρόπος υπολογισμού ενός πλάνου διανομής και τροχιών για την κατάσταση φόρτισης των μπαταριών μέσω της επίλυσης ενός προβλήματος βελτιστοποίησης. Στην επόμενη ενότητα, περιγράφεται η μέθοδος σχετικά με τον δίκαιο έλεγχο των μπαταριών που παρουσιάζεται στην εργασία [2] και στηρίζεται στις τροχιές. Τέλος, παρουσιάζονται δύο κριτήρια που μπορούν να χρησιμοποιηθούν για τη δίκαιη χρήση.

2.1 Κυρτή Βελτιστοποίηση

Ένα πρόβλημα κυρτής βελτιστοποίησης για ένα διάνυσμα $z \in \mathbb{R}^n$ έχει την παρακάτω μορφή:

$$\begin{aligned} & \text{minimize } f_0(z) \\ & \text{subject to } f_i(z) \leq b_i, i = 1, \dots, m \end{aligned}$$

Όπου οι συναρτήσεις $f_0, \dots, f_m : \mathbb{R}^n \rightarrow \mathbb{R}$ είναι κυρτές[12].

Η συνάρτηση f_i είναι κυρτή αν και μόνο αν για κάθε $z, v \in \mathbb{R}^n$ και για κάθε $a \in [0, 1]$ ισχύει $f_i(a \cdot z + (1 - a) \cdot v) \leq a \cdot f_i(z) + (1 - a) \cdot f_i(v)$.

Το διάνυσμα z αποτελεί το διάνυσμα των μεταβλητών που επιθυμείται η βελτιστοποίησή τους, οι οποίες ονομάζονται μεταβλητές απόφασης.

Για τα προβλήματα της οικογένειας αυτής υπάρχουν αλγόριθμοι που μπορούν να τα επιλύσουν αποδοτικά, γι' αυτό και τυποποίηση ενός προβλήματος ως πρόβλημα κυρτής βελτιστοποίησης ισοδυναμεί πρακτικά με επίλυση του προβλήματος. Αντίθετα, για τα προβλήματα μαθηματικής βελτιστοποίησης, όπου τουλάχιστον μία από τις συναρτήσεις δεν είναι κυρτή, τότε δεν υπάρχει γενική μέθοδος, η οποία να ελαχιστοποιεί αποδοτικά τη συνάρτηση.

Στην πράξη, συχνά, οι περιορισμοί του προβλήματος αφορούν ισότητες, δηλαδή έχουν τη μορφή $f_i(z) = b_i$. Τέτοιοι περιορισμοί ανάγονται στην παραπάνω μορφή και το πρόβλημα παραμένει κυρτό αν και μόνο αν η συνάρτηση $f_i(z)$ είναι γραμμική. Πράγματι, αν η συνάρτηση f_i είναι γραμμική, τότε η f_i και η $-f_i$ είναι κυρτές, οπότε ο περιορισμός ισότητας αντικαθίσταται από τους περιορισμούς $f_i(z) \leq b_i$ και $-f_i(z) \leq -b_i$, οι οποίοι ικανοποιούν τις

συνθήκες που ορίστηκαν παραπάνω. Υπενθυμίζεται ότι μία συνάρτηση είναι γραμμική αν έχει τη μορφή $f_i(z) = A \cdot z$, όπου A πραγματικός πίνακας $1 \times n$.

2.2 Συμβολισμοί

Για την περιγραφή των μεγεθών που παρουσιάζονται στο πρόβλημα βελτιστοποίησης στα έξυπνα δίκτυα ενέργειας που πραγματεύεται η παρούσα εργασία χρησιμοποιούνται οι παρακάτω συμβολισμοί:

- $P_{disp}(t)$: Η ισχύς στο PCC τη χρονική στιγμή t σύμφωνα με το πλάνο διανομής (dispatch plan)
- $P_{PCC}(t)$: Η παρατηρούμενη ισχύς στο PCC τη χρονική στιγμή t
- $L(t)$: Η συνολική ισχύς που παράγεται στο δίκτυο τη χρονική στιγμή t , χωρίς να λαμβάνονται υπόψιν οι μπαταρίες
- $B_i(t)$: Ενεργός ισχύς της μπαταρίας i τη χρονική στιγμή t
- $SoE_i(t)$: Η κατάσταση της ενέργειας της μπαταρίας i τη χρονική στιγμή t
- $x = (P_1, P_2, \dots, P_{N_B}, Q_1, Q_2, \dots, Q_{N_B})$: Το διάνυσμα των ενεργών και άεργων ισχύων όλων των μπαταριών για μία χρονική στιγμή
- u : Το διάνυσμα των ισχύων από τα μη ελέγξιμα στοιχεία του δικτύου μία χρονική στιγμή, ισχύει $sum(u) = L(t)$, εφόσον δεν υπάρχουν απώλειες
- $J_t(x|u)$: Το κόστος σχετικά με τον στόχο του δικτύου
- $w(t)$: Το βάρος που πολλαπλασιάζουμε το κόστος που αφορά τον συνολικό στόχο του δικτύου
- $C_{i,t}(B_i(t)|SoE_i(t))$: Το κόστος σχετικά με τον στόχο της μπαταρίας να ακολουθεί την προϋπολογισμένη τροχιά
- $B_i^*(SoE_i(t), t)$: Η προτίμηση για την ισχύ της μπαταρίας τη χρονική στιγμή t
- $r_i(t)$: Η απόκλιση της ισχύος της μπαταρίας i σε σχέση με την προτίμηση της τη χρονική στιγμή t , κανονικοποιημένη ως μέγιστη ισχύ της μπαταρίας
- N_B : πλήθος μπαταριών
- N_C : πλήθος συνδυαστικών περιορισμών
- SoE_i^{max} : Μέγιστη επιτρεπτή τιμή για το SoE της μπαταρίας i
- SoE_i^{min} : Ελάχιστη επιτρεπτή τιμή για το SoE της μπαταρίας i
- B_i^{max} : Μέγιστη επιτρεπτή τιμή για την ενεργό ισχύ της μπαταρίας i
- B_i^{min} : Ελάχιστη επιτρεπτή τιμή για την ενεργό ισχύ της μπαταρίας i

- SoE_j^{max,t_u} : Μέγιστη τιμή για το άθροισμα των SoE των μπαταριών που συμμετέχουν στον j -οστό περιορισμό που αφορά τη χρονική στιγμή t_u
- SoE_j^{min,t_u} : Ελάχιστη τιμή για το άθροισμα των SoE των μπαταριών που συμμετέχουν στον j -οστό περιορισμό που αφορά τη χρονική στιγμή t_u
- k : η αρχική χρονική στιγμή
- T : η τελική χρονική στιγμή
- Δt : χρόνος μεταξύ χρονικών στιγμών απόφασης

2.3 Υπολογισμός Τροχιών Κατάστασης Ενέργειας Μπαταριών και Πλάνου Διανομής

Η μέθοδος που παρουσιάζεται στο [10] προτείνει τον τρόπο υπολογισμού τροχιών για την κατάσταση ενέργειας κάθε μπαταρίας, καθώς και ενός πλάνου διανομής βάσει ενός συνόλου σεναρίων. Η μέθοδος αυτή επιλύει ένα πρόβλημα βέλτιστης ροής φορτίου (optimal power flow), το οποίο είναι μη κυρτό. Γι' αυτό και προτείνεται μία επαναληπτική μέθοδος, όπου σε κάθε επανάληψη απαιτείται η επίλυση ενός προβλήματος κυρτής βελτιστοποίησης. Στην παρούσα εργασία αγνοούνται οι απώλειες ενέργειας, επομένως το πρόβλημα βελτιστοποίησης είναι κυρτό και δεν απαιτούνται επαναλήψεις.

Ο υπολογισμός των τροχιών και του πλάνου διανομής προκύπτει από την επίλυση του εξής προβλήματος κυρτής βελτιστοποίησης:

$$\min_{B_1(t), \dots, B_{N_B}(t), P_{disp}(t)} \sum_{t=k}^{T-1} \left\{ \frac{1}{D} \sum_{d=1}^D [P_{PCC,d}(t) - P_{disp}(t)]^2 + w_c \cdot \sum_{i=1}^{N_B} |B_i(t)| \right\} \quad (2.1)$$

$$\text{subject to } P_{PCC,d}(t) = \sum_{i=1}^{N_B} B_i(t) + L_d(t) \quad (2.2)$$

$$SoE_i(t+1) = SoE_i(t) + B_i(t)\Delta t \quad (2.3)$$

$$B_i^{min} \leq B_i(t) \leq B_i^{max} \quad (2.4)$$

$$SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \quad (2.5)$$

$$\text{for } i \in \{1, 2, \dots, N_B\}, t \in \{k, \dots, T-1\}, j \in \{1, \dots, N_C\}$$

Ανάλογα με τους περαιτέρω στόχους που υπάρχουν μπορούν να προστεθούν κι άλλοι περιορισμοί στο πρόβλημα, αρκεί το σύνολο των εφικτών λύσεων να παραμένει κυρτό. Για παράδειγμα, μπορεί να προστεθούν περιορισμοί σχετικά με τα συνολικά αποθέματα (reserves) την τελική χρονική στιγμή, δηλαδή $SoE^{min,T} \leq \sum_{i=1}^{N_B} SoE_i(T) \leq SoE^{max,T}$

Ο αριθμός D υποδεικνύει το πλήθος των σεναρίων που χρησιμοποιήθηκαν για τον υπολογισμό των τροχιών, ενώ με $P_{PCC,d}(t)$ συμβολίζεται η ισχύς στο PCC τη χρονική στιγμή t για το d -οστό σενάριο.

Ο όρος $\sum_{i=1}^{N_B} |B_i(t)|$ ονομάζεται battery cycling cost. Ο όρος αυτός αποθαρρύνει την κυκλική φόρτιση-εκφόρτιση, όμως δεν είναι αρκετός για τη δίκαιη χρήση μπαταριών όπως αποδεικνύεται στη συνέχεια. Η βαρύτητα που δίνεται στο κόστος αυτό καθορίζεται από την

τιμή του w_c . Συνήθως η τιμή αυτή είναι μικρή, ώστε να μην επηρεάζει σημαντικά το κόστος του κύριου στόχου που είναι να ακολουθείται το πλάνο διανομής.

Η παραπάνω μέθοδος μπορεί να χρησιμοποιηθεί και για τον υπολογισμό τροχιών για το SoE κάθε μπαταρίας, ακόμη και στην περίπτωση που το πλάνο διανομής είναι γνωστό και υπολογισμένο εξ αρχής (έστω ότι δίνεται από την ακολουθία $P_{DP}(t)$). Στην περίπτωση αυτή αρκεί να προστεθούν στο πρόβλημα βελτιστοποίησης που περιγράφηκε οι περιορισμοί:

$$P_{disp}(t) = P_{DP}(t), \forall t \in \{k, \dots, T - 1\} \quad (2.6)$$

2.4 Δίκαιος Έλεγχος Μπαταριών

Στην ενότητα αυτή παρουσιάζεται συνοπτικά η μέθοδος που χρησιμοποιείται στο [2], για τη δίκαιη χρήση των μπαταριών. Στην ενότητα 4.4 χρησιμοποιείται αυτή η μέθοδος για την επίλυση του προβλήματος που πραγματεύεται η παρούσα εργασία, παρά τις διαφορές που τονίζονται ότι υπάρχουν στα δύο προβλήματα. Στο έργο αυτό, ο έλεγχος πραγματοποιείται μέσω της κεντριοποιημένης επίλυσης ενός προβλήματος κυρτής βελτιστοποίησης για κάθε χρονική στιγμή ανεξάρτητα. Τα κύρια αποτελέσματα που προκύπτουν από την εργασία είναι ο σχεδιασμός της συνάρτησης κόστους που εξασφαλίζει τη δικαιοσύνη σε σχέση με τον στόχο κάθε μπαταρίας να ακολουθεί την προϋπολογισμένη τροχιά της, καθώς και η εύρεση τύπου για τον υπολογισμό του συντελεστή που εξισορροπεί τη βαρύτητα που δίνεται στον στόχο αυτό σε σχέση με τον συνολικό στόχο του δικτύου.

Θεωρώντας ένα δίκτυο ενέργειας, το οποίο αποτελείται από N_B μπαταρίες, οι οποίες μπορούν να προσφέρουν ή απορροφούν ενεργό και άεργο ισχύ $x_i = (P_i, Q_i) \in A_i, i = \{1, \dots, N_B\}$, όπου $A_i = [P_i^{min}, P_i^{max}] \times [Q_i^{min}, Q_i^{max}]$ το σύνολο των εφικτών τιμών για κάθε μπαταρία λόγω των περιορισμών. Επειδή η εργασία [2] χρησιμοποιεί παραδοσιακές τεχνικές βελτιστοποίησης, η συνάρτηση σχεδιάζεται, ώστε να είναι κυρτή, θεωρώντας μόνο τους απαραίτητους περιορισμούς, προκειμένου το πρόβλημα να επιλύεται αποδοτικά, ώστε ο έλεγχος να πραγματοποιείται σε πραγματικό χρόνο.

Για κάθε μπαταρία ορίζεται μία συνάρτηση κόστους $C_i(x_i|S_i, SoE_i)$, η οποία θα αναφέρεται ως στόχος μπαταρίας (battery objective) (S_i είναι η ονομαστική φαινόμενη ισχύς του μετατροπέα της μπαταρίας). Η συνάρτηση αυτή δηλώνει την προτίμηση της μπαταρίας προς συγκεκριμένο σημείο από το σύνολο των εφικτών αποφάσεων. Εκτός του ατομικού στόχου κάθε μπαταρίας, υπάρχει και ένας γενικός στόχος του δικτύου (grid objective), για τον οποίο ορίζεται μία συνάρτηση κόστους $J_t(x|u)$, όπου με u συμβολίζεται η ισχύς που προκύπτει από τα μη ελεγχόμενα στοιχεία του δικτύου. Ο στόχος του δικτύου μπορεί να αφορά την ισχύ στο PCC ή τον έλεγχο της τάσης ή του ρεύματος σε συγκεκριμένα σημεία του δικτύου.

Έτσι, ο έλεγχος πραγματοποιείται μέσω επίλυσης του ακόλουθου προβλήματος βελτιστο-

ποίησης, ανεξάρτητα για κάθε χρονική στιγμή:

$$\min_{x_1, \dots, x_{N_B}} C_t(x|u) \quad (2.7)$$

$$\text{subject to } x_i \in A_i, \forall i = \{1, \dots, N_B\} \quad (2.8)$$

$$C_t(x|u) = \sum_{i=1}^{N_B} C_i(x_i|S_i, SoE_i) + J(x|u) \quad (2.9)$$

Σχετικά με την παραγωγή u των μη ελεγχόμενων στοιχείων, καθώς δεν είναι γνωστή εκείνη τη χρονική στιγμή λόγω της στοχαστικότητας και των αβεβαιοτήτων χρησιμοποιείται μία πρόβλεψη αυτής της τιμής (π.χ. η τιμή τους την προηγούμενη χρονική στιγμή).

Η επίλυση του παραπάνω προβλήματος μπορεί να πραγματοποιηθεί αποδοτικά, χρησιμοποιώντας παραδοσιακές τεχνικές βελτιστοποίησης, εφόσον η συνάρτηση κόστους είναι κυρτή. Στη συνέχεια της ενότητας παρουσιάζεται η μέθοδος σχεδιασμού των συναρτήσεων κόστους μπαταρίας και δικτύου, ώστε η συνολική συνάρτηση να είναι κυρτή.

Σχετικά με τη συνάρτηση κόστους μπαταρίας, αρχικά υπολογίζεται ένα σημείο, το οποίο ονομάζεται προτίμηση της μπαταρίας. Το σημείο αυτό προκύπτει βάσει της τροχιάς της μπαταρίας και της τιμής του SoE εκείνη τη χρονική στιγμή. Έπειτα, ορίζεται ως battery objective μία κυρτή συνάρτηση για την απόκλιση από το σημείο προτίμησης. Η συνάρτηση αναφορικά με το grid objective μπορεί να σχεδιαστεί και αυτή να είναι κυρτή, όπως παρουσιάζεται στη συνέχεια. Προκειμένου να εξισορροπηθεί η βαρύτητα που δίνεται στους δύο στόχους, παρουσιάζεται αναλυτικός τύπος για τον υπολογισμό της κατάλληλης τιμής για τα βάρη. Η συνολική μέθοδος που χρησιμοποιείται για τον έλεγχο στην εργασία σε μορφή αλγορίθμου είναι η ακόλουθη, με τις συναρτήσεις που αναφέρονται να ορίζονται στις επόμενες υποενότητες:

ΑΛΓΟΡΙΘΜΟΣ 2.1: Δίκαιος έλεγχος μπαταριών μέσω κυρτής βελτιστοποίησης

```

for ( $t \leftarrow k \cdot t < T \cdot + + t$ ) do
   $w \leftarrow \text{ComputeWeight}(u)$ 
  for all batteries do
     $x_i^* \leftarrow \text{ComputePreference}(SoE_i, t)$ 
  end for
   $x_1, x_2, \dots, x_{N_B} \leftarrow \text{SolveOptimizationProblem}((2.7)-(2.9))$ 
  for all batteries do
     $SoE_i \leftarrow \text{UpdateSoE}(SoE_i, x_i)$ 
  end for
   $u \leftarrow \text{UpdateEstimation}(u(t))$ 
end for

```

2.4.1 Συνάρτηση κόστους μπαταρίας

Ο ατομικός στόχος της μπαταρίας είναι να ικανοποιήσει την προτίμηση της. Ως προτίμηση ορίζεται ένα σημείο $x_i^*(SoE_i, t) = (P_i^*(SoE_i, t), Q_i^*(SoE_i, t)) \in A_i$, το οποίο μπορεί να οριστεί ως το σημείο που φέρνει το SoE_i όσο πιο κοντά στην τροχιά που θέλει να ακολουθήσει. Σε περίπτωση που δεν υπάρχει κάποιος αντικρουόμενος στόχος για το δίκτυο, τότε η μπαταρία

πρέπει να εκτελεί την προτίμηση της.

Κατά τον σχεδιασμό της συνάρτησης κόστους πρέπει να ληφθεί υπόψιν και η δίκαιη χρήση μεταξύ των μπαταριών. Αν (P_i^{opt}, Q_i^{opt}) είναι το σημείο που ελαχιστοποιείται το συνολικό κόστος, τότε ορίζονται ως λόγος προτίμησης οι παραστάσεις $r_{P,i} = \frac{P_i^{opt} - P_i^*}{S_i}$ και $r_{Q,i} = \frac{Q_i^{opt} - Q_i^*}{S_i}$. Βάσει αυτού, θεωρείται δίκαιο η τιμή του λόγου προτίμησης να είναι ανάλογη της επιρροής που έχει η αντίστοιχη μπαταρία στον στόχο του δικτύου. Δηλαδή να ισχύει για κάθε μπαταρία:

$$r_{P,i} = \beta_P \frac{\partial J(x^{opt}|u)}{\partial P_i}, r_{Q,i} = \beta_Q \frac{\partial J(x^{opt}|u)}{\partial Q_i}, \forall i = \{1, \dots, N_B\} \quad (2.10)$$

Με αυτόν τον τρόπο οι μπαταρίες με μεγάλη επιρροή μπορούν να εκμεταλλεύονται για να επιτευχθεί ο συνολικός στόχος ακόμη κι αν αυτός δεν συμβαδίζει με την προτίμησή τους.

Ο στόχος της συνάρτησης κόστους μπαταρίας είναι να ικανοποιείται η προτίμηση της μπαταρίας και να μειωθούν οι εσωτερικές της απώλειες. Έτσι, η συνάρτηση αυτή πρέπει να ελαχιστοποιείται στο σημείο προτίμησης της και να προστίθεται ποινή ανάλογη του τετραγώνου της ενεργού και της άεργου ισχύος. Βάσει αυτών η συνάρτηση κόστους, έχει τη γενική μορφή:

$$C_i(P_i, Q_i|S_i, SoE_i) = a_P \cdot \left(\frac{P_i^2}{4S_i} - \frac{P_i^*(SoE_i, t)}{2S_i} P_i \right) + a_Q \cdot \left(\frac{Q_i^2}{4S_i} - \frac{Q_i^*(SoE_i, t)}{2S_i} Q_i \right) + c \quad (2.11)$$

Τα a_P, a_Q καθορίζουν τη βαρύτητα που δίνεται σε κάθε όρο και πέρνουν θετικές τιμές, ώστε η συνάρτηση να είναι αυστηρά κυρτή. Με c συμβολίζεται μία σταθερά, καθώς η πρόσθεση οποιουδήποτε σταθερού αριθμού δεν επηρεάζει το ελάχιστο της συνάρτησης. Σημειώνεται ότι λόγω της δέσμευσης στις τιμές των S_i, SoE_i η σταθερά c μπορεί να θεωρηθεί ως συνάρτηση αυτών των μεγεθών.

Σχετικά με τη δίκαιη χρήση, αν δύο μπαταρίες i, j έχουν την ίδια προτίμηση, κανονικοποιημένη ως προς την ονομαστική ισχύ, δηλαδή $\frac{x_i^*(SoE_i, t)}{S_i} = \frac{x_j^*(SoE_j, t)}{S_j}$ και αντίστοιχα τους ίδιους περιορισμούς, δηλαδή $\frac{P_i}{S_i} \in [(\frac{P}{S})_{min}, (\frac{P}{S})_{max}]$, $\frac{Q_i}{S_i} \in [(\frac{Q}{S})_{min}, (\frac{Q}{S})_{max}]$ και αντίστοιχα $\frac{P_j}{S_j} \in [(\frac{P}{S})_{min}, (\frac{P}{S})_{max}]$, $\frac{Q_j}{S_j} \in [(\frac{Q}{S})_{min}, (\frac{Q}{S})_{max}]$, τότε η λύση του προβλήματος βελτιστοποίησης (2.7) ικανοποιεί της παρακάτω συνθήκες:

$$|r_{P,i} - r_{P,j}| \leq \frac{2}{a_P} \left| \frac{\partial J(x^{opt}|u)}{\partial P_i} - \frac{\partial J(x^{opt}|u)}{\partial P_j} \right| \quad (2.12)$$

$$|r_{Q,i} - r_{Q,j}| \leq \frac{2}{a_Q} \left| \frac{\partial J(x^{opt}|u)}{\partial Q_i} - \frac{\partial J(x^{opt}|u)}{\partial Q_j} \right| \quad (2.13)$$

Δηλαδή η διαφορά μεταξύ των λόγων προτίμησης των δύο μπαταριών είναι το πολύ ίση με αυτήν της περίπτωση της δίκαιης χρήσης, όπως ορίστηκε προηγουμένως.

Σχετικά με την επιλογή της τιμής των $x^*(SoE_i, t)$, $x \in \{P, Q\}$ θεωρώντας γνωστή μια τροχιά SoE_i^* για κάθε μπαταρία, η προτίμηση για κάθε χρονική στιγμή υπολογίζεται μέσω της επόμενης συνάρτησης, η οποία αντιστοιχεί στη συνάρτηση ComputePreference του αλγορίθμου:

$$x^*(SoE_i, t) = \begin{cases} x_i^{max} & , \Delta SoE_i \leq -\epsilon_i \\ -\frac{\Delta SoE_i \cdot x_i^{max}}{\epsilon_i} & , |\Delta SoE_i| < \epsilon_i \\ -x_i^{max} & , \Delta SoE_i \geq \epsilon_i \end{cases} \quad (2.14)$$

Όπου $\Delta SoE_i = SoE_i(t) - SoE_i^*(t+1)$ και $\epsilon_i \geq x_i^{max} \cdot \Delta t$, όπου $x_i^{max} \in \{P_i^{max}, Q_i^{max}\}$ αντίστοιχα με το x , παράμετρος που καθορίζει πόσο κοντά στην τροχιά επιθυμεί να φτάσει η μπαταρία σύμφωνα με την προτίμηση της. Η τιμή $\epsilon_i = x_i^{max} \cdot \Delta t$ καθορίζει την προτίμηση ως την ισχύ που οδηγεί την κατάσταση ενέργειας της μπαταρίας στην πλησιέστερη δυνατή τιμή στην κατάσταση που καθορίζεται από την τροχιά της. Μικρότερες τιμές από την παραπάνω πρέπει να αποφεύγονται, γιατί υπάρχει ο κίνδυνος να γίνονται ταλαντώσεις, δηλαδή να αλλάζει συνεχώς πρόσημο η προτίμηση κατά τη διάρκεια μιας ακολουθίας λήψης αποφάσεων.

2.4.2 Συνάρτηση κόστους δικτύου- Υπολογισμός βαρών

Η συνάρτηση κόστους δικτύου (grid objective) αντιπροσωπεύει τον κοινό στόχο του δικτύου. Ο στόχος είναι η βελτιστοποίηση ενός συνόλου μεταβλητών του δικτύου $y = (y_1, \dots, y_M)$, όπου y_m μπορεί να είναι η έγχυση ενεργού ή άεργου ισχύος σε έναν κόμβο, η τάση σε συγκεκριμένο κόμβο ή το ρεύμα ενός κλάδου. Επομένως, η συνάρτηση έχει γενικά τη μορφή:

$$J(x|u) = \sum_{m=1}^M \omega_m J_m(y_m(x|u)) \quad (2.15)$$

Όπου J_m η συνάρτηση, που δίνει το κόστος για την τιμή της μεταβλητής y_m και $\omega_m > 0$ το αντίστοιχο βάρος.

Οι μεταβλητές του δικτύου συνδεούνται μεταξύ τους και με την ισχύ των μπαταριών και των μη ελεγχόμενων στοιχείων, μέσω των εξισώσεων ροής φορτίου. Αν V_b η τάση στον ζυγό b και $P_b - jQ_b$ η έγχυση ισχύος στον ίδιο ζυγό, τότε αυτά τα μεγέθη συνδέονται μέσω της σχέσης:

$$P_b - jQ_b = V_b^* \cdot \sum_{l=1}^n (Y_{bl} \cdot V_l)$$

Όπου Y_{bl} η (μιγαδική) αγωγιμότητα μεταξύ των ζυγών b, l . Λόγω των εξισώσεων ροής φορτίου, οι μεταβλητές αυτές δεν μπορούν να εκφραστούν ως κυρτές συναρτήσεις των ενεργών και άεργων ισχύων (x, u) . Έτσι, για να είναι σχεδιαστεί μία κυρτή συνάρτηση κόστους χρησιμοποιείται η γραμμικοποίηση αυτών των μεγεθών ως εξής:

$$\begin{aligned} y_m(x|u) = & y_m^0 + \sum_{i=1}^{N_B} (K_{P,i}^m (P_i - P_i^0) + K_{Q,i}^m (Q_i - Q_i^0)) \\ & + \sum_{n=1}^U (K_{P,n}^m (P_n - P_n^0) + K_{Q,n}^m (Q_n - Q_n^0)) \end{aligned} \quad (2.16)$$

Όπου $y_m^0, P_i^0, Q_i^0, P_n^0, Q_n^0$ δηλώνουν την τιμή των αντίστοιχων μεγεθών y_m, P_i, Q_i, P_n, Q_n την προηγούμενη χρονική στιγμή, ενώ τα $K_{P,i}^m, K_{Q,i}^m, K_{P,n}^m, K_{Q,n}^m$ ονομάζονται συντελεστές ευαισθησίας, και πρόκειται ουσιαστικά για τις μερικές παραγώγους της μεταβλητής y_m ως προς τα αντίστοιχα μεγέθη υπολογισμένες στα σημεία των τιμών της προηγούμενης χρονικής στιγμής. Ο δείκτης i χρησιμοποιείται για τις μπαταρίες, ενώ με n δηλώνονται τα μη ελέγξιμα στοιχεία του δικτύου.

Ακόμη, για τη συνάρτηση J_m υποτίθεται ότι έχει τη μορφή

$$J_m(y_m) = J_m^*(y_m - y_m^*) \quad (2.17)$$

Όπου $J_m^*(y)$ είναι μία αυστηρά κυρτή και άρτια συνάρτηση του y , ενώ y_m^* είναι το σημείο που ελαχιστοποιείται η συνάρτηση $J_m(y_m)$. Αντίστοιχα, ορίζεται ως y_m^{inf} το σημείο ελαχίστου της συνάρτησης $J_m(y_m)$ στο πεδίο τιμών της y_m , σύμφωνα με τους περιορισμούς του προβλήματος.

Για τον θεωρητικό υπολογισμό των βαρών ορίζονται οι παρακάτω βοηθητικές συναρτήσεις:

$$K_t(w_m, d) = \frac{1}{w_m J_m'(y_m^* + d)} \quad (2.18)$$

$$h(K) = |y_m^* - y_m^{inf}| + \sum_{|K_{x,i}^m| < K} |K_{x,i}^m| |x_i^{max} - x_i^{min}| \quad (2.19)$$

$$g(w_m, d) = \max\{d, h(K_t(w_m, d))\} \quad (2.20)$$

$$f(w_m) = \min_{d>0} g(w_m, d) \quad (2.21)$$

Θεωρώντας ότι υπάρχει ένας στόχος $d_{max,m}^*$ για τη μέγιστη απόκλιση της τιμής της μεταβλητής y_m στο σημείο y_m^{opt} της βέλτιστης λύσης του προβλήματος (2.7) από την ελάχιστη τιμή του κόστους του δικτύου y_m^* , δηλαδή τα βάρη πρέπει να πάρουν τιμή τέτοια, ώστε $|y_m^{opt} - y_m^*| \leq d_{max,m}^*$. Το ακόλουθο λήμμα δίνει ένα άνω φράγμα για τη διαφορά αυτή.

Λήμμα

Αν το κόστος δικτύου δίνεται από συνάρτηση $J(x|u) = J_m(y_m(x|u))$, όπου $J_m(y_m)$ είναι αυστηρά κυρτή συνάρτηση της μεταβλητής y_m , η οποία έχει τη μορφή (2.16). Θεωρώντας ότι οι μπαταρίες έχουν περιορισμούς της μορφής $x_i \in [x_i^{min}, x_i^{max}]$, όπου $x_i = \{P_i, Q_i\}$, τότε η συνάρτηση $f(w_m)$ είναι άνω φράγμα της $|y_m^{opt} - y_m^*|$.

Στην πράξη υπάρχει και η αβεβαιότητα για την ισχύ των μη ελέγξιμων στοιχείων. Έτσι, η πραγματική τιμή της μεταβλητής θα είναι $y_m^{real} = y_m^{opt} + \Delta y_m$, όπου $\Delta y_m = \sum_{n=1}^U (K_{P,n}^m \Delta P_n + K_{Q,n}^m \Delta Q_n)$ με $\Delta P_n, \Delta Q_n$ να συμβολίζουν τη διαφορά της πρόβλεψης με την πραγματική τιμή της ενεργού και της άεργου ισχύος αντίστοιχα. Συνδυάζοντας αυτό με το προηγούμενο λήμμα προκύπτει το παρακάτω θεώρημα.

Θεώρημα

Εφόσον ισχύουν οι υποθέσεις του προηγούμενου λήμματος, τότε

$$|y_m^{real} - y_m^*| \leq f(w_m) + |\Delta y_m|_{max}$$

Όπου $|\Delta y_m|_{max} = \max_{P_n, Q_n} |\Delta y_m|$

Το προηγούμενο άνω φράγμα ορίζεται ως η συνάρτηση $f_u(w_m) = f(w_m) + |\Delta y_m|_{max}$. Η συνάρτηση αυτή έχει τις παρακάτω ιδιότητες

- Είναι φθίνουσα ως προς w_m
- Έχει άνω φράγμα $f_{UB} = |y_m^* - y_m^{inf}| + \sum_{i \in I} |K_{x,i}^m| |x_i^{max} - x_i^{min}| + |\Delta y_m|_{max}$

- Έχει κάτω φράγμα $f_{LB} = |y_m^* - y_m^{inf}| + |\Delta y_m|_{max}$
- Αν $y_m^* \neq y_m^{inf}$, τότε $f_u(w_m) = f_{LB}, \forall w_m \geq \frac{1}{|K^m|_{min}|J'_m(y_m^{inf})|}$, όπου $|K^m|_{min} = \min |K_{x,i}^m|$

Το επιθυμητό βάρος είναι αυτό που επιτυγχάνει τον στόχο του δικτύου, ενώ εξισορροπεί τη βαρύτητα του στόχου κάθε μπαταρίας. Έτσι, πρόκειται για το ελάχιστο $w_m > 0$, για το οποίο ισχύει $f_u(w_m) \leq d_{max,m}^*$. Χρησιμοποιώντας όλα τα παραπάνω, στο [2] καταλήγουν στον παρακάτω αναλυτικό τύπο για τον υπολογισμό των βαρών, ο οποίος και αντιστοιχεί στη συνάρτηση $\text{ComputeWeight}(u)$ του αλγορίθμου.

$$w_m = \begin{cases} 0 & , d_{max,m}^* > f_{UB} \\ \frac{1}{|K^m|_{min}|J'_m(y_m^{inf})|} & , d_{max,m}^* < f_{LB} \\ \frac{1}{K_t^m|J'_m(y_m^* + d_{max,m}^* - |\Delta y_m|_{max})|} & , \text{else} \end{cases} \quad (2.22)$$

Όπου

- $|K^m|_{min}$ ο ελάχιστος συντελεστής ευαισθησίας στην εξίσωση (2.16)
- K_t^m είναι το μέγιστο K τέτοιο, ώστε $h(K) < d_{max,m}^*$

Επίσης, για τον υπολογισμό των βαρών με βάση τον παραπάνω τύπο έγινε η υπόθεση $a_p \leq 1, a_Q \leq 1$ για τους συντελεστές της εξίσωσης (2.11).

Η ερμηνεία των βαρών είναι η ακόλουθη. Η τιμή του f_{UB} είναι το άθροισμα της απόκλισης που υπάρχει σίγουρα λόγω των περιορισμών, της μέγιστης απόκλισης που μπορεί να υπάρχει από τη βέλτιστη λύση, δηλαδή να παρέχουν τη μέγιστη ισχύ, ενώ πρέπει την ελάχιστη ή αντίστροφα, και την αβεβαιότητα. Αν ο στόχος έχει μεγαλύτερη τιμή από αυτό το άνω φράγμα, τότε ικανοποιείται από οποιαδήποτε λύση, οπότε το βάρος παίρνει την τιμή 0, ώστε κάθε μπαταρία να ικανοποιεί την προτίμησή της. Αντίστοιχα, η τιμή του f_{LB} είναι η τιμή της απόκλισης από τη βέλτιστη λύση λόγω των περιορισμών συν την αβεβαιότητα. Αυτό σημαίνει ότι οποιοσδήποτε στόχος μικρότερος από αυτή την τιμή είναι ανέφικτος. Τότε θα έπρεπε να δοθεί το μέγιστο δυνατό βάρος, ώστε η λύση να είναι όσο πιο κοντά στον στόχο. Όμως, από την τέταρτη ιδιότητα της $f_u(w_m)$. Για τις ενδιάμεσες περιπτώσεις η τιμή προκύπτει από το πρόβλημα βελτιστοποίησης, που περιγράφηκε. Αφού η J_m είναι αυστηρά κυρτή, η J'_m είναι γνησίως αύξουσα, οπότε για μεγαλύτερα d_{max}^m το βάρος παίρνει μικρότερη τιμή, όπως αναμένεται και διαισθητικά.

2.5 Κριτήρια Δικαιοσύνης

Στην παρούσα ενότητα παρουσιάζονται δύο κριτήρια βάσει των οποίων μπορεί να χαρακτηριστεί δίκαιη η χρησιμοποίηση των μπαταριών. Για κάθε κριτήριο παρουσιάζεται διαισθητικά ο λόγος που παρέχει δικαιοσύνη, καθώς και τα πλεονεκτήματά του.

2.5.1 Δικαιοσύνη βάσει των περιορισμών ισχύος

Θεωρώντας ότι δύο μπαταρίες έχουν κοινό στόχο, όπως να ακολουθείται ένα πλάνο διανομής ή να ικανοποιείται ένας συνολικός περιορισμός για τα αποθέματα ενέργειας, θα ήταν

άδικο η μία να φορτίζει και η άλλη να ξεφορτίζει την ίδια χρονική στιγμή. Αντίστοιχα, άδικο θα ήταν να χρησιμοποιούταν μόνο η μία από τις δύο και η άλλη να παρέμενε ανενεργή. Ένας τρόπος για να αποφευχθούν τα παραπάνω είναι η χρησιμοποίηση κάθε μπαταρίας να είναι ανάλογη της μέγιστης δυνατής ενέργειας της.

Έστω ότι μία δεδομένη χρονική στιγμή t δύο μπαταρίες i, j οφείλουν να φορτίσουν για να ικανοποιήσουν τους στόχους του δικτύου. Τότε, το κριτήριο δικαιοσύνης που περιγράφεται στην προηγούμενη παράγραφο απαιτεί να ικανοποιούν οι μπαταρίες τη σχέση

$$\frac{B_i(t)}{B_i^{max}} = \frac{B_j(t)}{B_j^{max}} \quad (2.23)$$

Είναι αντιληπτό από την παραπάνω σχέση ότι αν δεν χρησιμοποιείται η μία μπαταρία, τότε δεν θα χρησιμοποιείται ούτε η άλλη, ενώ μία μπαταρία επιλέγει τη μέγιστη δυνατή φόρτιση, αν και μόνο αν ισχύει το ίδιο και για την άλλη.

Αντίστοιχα, για μία χρονική στιγμή t κατά την οποία οι μπαταρίες οφείλουν να εκφορτίσουν, τότε το κριτήριο δικαιοσύνης εκφράζεται από τη σχέση

$$\frac{B_i(t)}{B_i^{min}} = \frac{B_j(t)}{B_j^{min}} \quad (2.24)$$

Είναι συνηθισμένο σε πολλές περιπτώσεις να ισχύει για τους περιορισμούς η σχέση $B_i^{min} = -B_i^{max}$. Σε αυτή την περίπτωση μπορεί να χρησιμοποιηθεί για κάθε περίπτωση μία από τις παραπάνω σχέσεις (π.χ. η (2.23)). Μάλιστα, είναι εμφανές ότι το πρόσημο των δράσεων πρέπει να είναι ίδιο, για να ικανοποιείται η σχέση, οπότε θα συμφωνούν ως προς τη δράση τους. Αυτό ισχύει γενικότερα για κάθε περίπτωση όπου ο λόγος $\frac{B_{min}}{B_{max}}$ είναι ο ίδιος για όλες τις μπαταρίες. Αυτό αποτελεί ένα σημαντικό πλεονέκτημα του κριτηρίου, καθώς μπορεί να έχει πρακτικές εφαρμογές σε περιπτώσεις, όπου δεν είναι γνωστό εκ των προτέρων, αν πρέπει να γίνει φόρτιση ή εκφόρτιση.

2.5.2 Δικαιοσύνη βάσει της τρέχουσας κατάστασης της ενέργειας

Η έννοια της δικαιοσύνης, σχετίζεται με τον στόχο που υπάρχει, οπότε ένα κριτήριο δεν μπορεί να θεωρείται το κατάλληλο σε κάθε περίπτωση. Για παράδειγμα, έστω 2 μπαταρίες i, j με $B_i^{max} > B_j^{max}$, όπου το SoE της μπαταρίας είναι αρκετά κοντά στη μέγιστη τιμή του SoE_i^{max} σε αντίθεση με αυτό της μπαταρίας j . Έστω ότι το φορτίο είναι τέτοιο, ώστε οι μπαταρίες να πρέπει να φορτίσουν, ώστε να ικανοποιήσουν τον στόχο τους. Χρησιμοποιώντας το προηγούμενο κριτήριο θα είχε αποτέλεσμα η μπαταρία i να έφτανε τη μέγιστη τιμή για το SoE σύντομα και στη συνέχεια να μην μπορεί να χρησιμοποιηθεί. Σε τέτοιες περιπτώσεις το ακόλουθο κριτήριο θα ήταν πιο κατάλληλο.

Έστω μία χρονική στιγμή t όπου οι μπαταρίες i, j πρέπει να φορτίσουν. Τότε η δίκαιη χρήση επιτυγχάνεται, αν ικανοποιείται η σχέση

$$\frac{B_i(t)}{SoE_i^{max} - SoE_i(t)} = \frac{B_j(t)}{SoE_j^{max} - SoE_j(t)} \quad (2.25)$$

Η σχέση αυτή ενθαρρύνει την μπαταρία που απέχει περισσότερο από το μέγιστο να φορτίσει

περισσότερο. Αντίστοιχα, αν μία χρονική στιγμή οι μπαταρίες πρέπει να εκφορτίσουν, τότε η σχέση δικαιοσύνης είναι η

$$\frac{B_i(t)}{SoE_i(t) - SoE_i^{min}} = \frac{B_j(t)}{SoE_j(t) - SoE_j^{min}} \quad (2.26)$$

Το κριτήριο αυτό επιτυγχάνει τη σύγκλιση των καταστάσεων της ενέργειας μεταξύ των μπαταριών [11].

Το κύριο μειονέκτημα αυτής της μεθόδου είναι ότι απαιτείται να είναι γνωστό για κάθε χρονική στιγμή αν πρέπει να γίνει φόρτιση ή εκφόρτιση, με αποτέλεσμα μην είναι άμεσα χρησιμοποιήσιμο σε περιπτώσεις όπου οι αποφάσεις λαμβάνονται αποκεντρωμένα, σε πραγματικό χρόνο και χωρίς επικοινωνία. Ακόμη, αν απαιτείται να ικανοποιούνται αυστηρά οι παραπάνω σχέσεις, τότε δεν λαμβάνονται καθόλου υπόψιν οι μέγιστες δυνατότητες του δικτύου, οπότε μπορεί να οδηγήσουν σε υποβέλτιστα αποτελέσματα.

Κεφάλαιο 3

Μηχανική Μάθηση

Στο κεφάλαιο αυτό συγκεντρώνεται η θεωρία σχετικά με το πεδίο της μηχανικής μάθησης, όπως χρησιμοποιείται στην παρούσα εργασία. Αρχικά, παρουσιάζεται μια εισαγωγή στο ευρύτερο πεδίο της μηχανικής μάθησης. Έπειτα, περιγράφεται η απαραίτητη θεωρία σχετικά με τα νευρωνικά δίκτυα. Η τρίτη ενότητα περιγράφει την αρχιτεκτονική των βαθιών νευρωνικών δικτύων, ενώ καταπιάνεται και με θέματα σχετικά με την εκπαίδευσή τους. Στη συνέχεια, παρουσιάζονται οι έννοιες των Μαρκοβιανών Αλυσίδων Απόφασης (Markov Decision Process - MDP) και των Μαρκοβιανών Παιγνίων (Markov Games), που αποτελούν θεμελιώδεις τρόπους αναπαράστασης του περιβάλλοντος στην ενισχυτική μάθηση. Στο πέμπτο κεφάλαιο, ορίζεται η ενισχυτική μάθηση και περιγράφονται κάποιοι από τους βασικούς αλγόριθμους του πεδίου. Τέλος, παρουσιάζεται το θεωρητικό υπόβαθρο της ενισχυτικής μάθησης πολλαπλών πρακτόρων (Multi-Agent Reinforcement Learning) με εκτενή αναφορά στον αλγόριθμο Multi-Agent Deep Deterministic Policy Gradient (MADDPG), ο οποίος και υλοποιείται στην τρέχουσα εργασία.

3.1 Μηχανική Μάθηση

Η μηχανική μάθηση (machine learning) αποτελεί ένα από τα σημαντικότερα πεδία του ευρύτερου τομέα της τεχνητής νοημοσύνης (artificial intelligence). Τα εξαιρετικά αποτελέσματα που επιτυγχάνει σε πληθώρα προβλημάτων την έχουν καταστήσει ως το κατ' εξοχήν πεδίο έρευνας στο συγκεκριμένο τομέα.

3.1.1 Ορισμός

Στο [13] δίνεται ο ακόλουθος ορισμός για τη μηχανική μάθηση:

“Ένα πρόγραμμα υπολογιστεί μαθαίνει από την εμπειρία E , σχετικά με ένα σύνολο εργασιών T και μία μετρική επίδοσης P , αν η επίδοση στις εργασίες που ανήκουν στο T , σύμφωνα με τη μετρική P , βελτιώνεται μέσω της εμπειρίας E .”

Σκοπός της μηχανικής μάθησης είναι η εκμάθηση λύσεων χωρίς οι υπολογιστές να προγραμματίζονται ρητά σε ένα συγκεκριμένο πρόβλημα. Οι λόγοι που οδηγούν σε μία τέτοια προσέγγιση είναι η αδυναμία πρόβλεψης όλων των πιθανών καταστάσεων που μπορεί να βρεθεί ένα πρόγραμμα εκ των προτέρων, καθώς και η έλλειψη γνωστών λύσεων σε προβλήματα που είναι εύκολα για έναν άνθρωπο, άλλα δύσκολα για έναν υπολογιστή, όπως είναι αυτό της αναγνώρισης προσώπου[14].

3.1.2 Κατηγορίες Μηχανικής Μάθησης

Το αντικείμενο της μηχανικής μάθησης διαιρείται σε τρεις κύριες κατηγορίες, με βάση τον τρόπο που γίνεται η αξιολόγηση της εμπειρίας εκπαίδευσης: την επιβλεπόμενη μάθηση (supervised learning), τη μη επιβλεπόμενη μάθηση (unsupervised learning) και την ενισχυτική μάθηση (reinforcement learning) [14]. Οι κατηγορίες αυτές διαφέρουν τόσο στην γνώση που απαιτείται σχετικά με τα δεδομένα, όσο και στον τελικό τους στόχο. Ακολουθούν τα κυριότερα γνωρίσματα κάθε κατηγορίας, καθώς και προβλήματα όπου βρίσκουν εφαρμογή.

Επιβλεπόμενη Μάθηση

Στην περίπτωση της επιβλεπόμενης μάθησης το σύνολο των δεδομένων εκπαίδευσης περιγράφεται από το ζεύγος της εισόδου, δηλαδή της περιγραφής των χαρακτηριστικών, και της εξόδου, η οποία ονομάζεται ετικέτα (label). Σκοπός της είναι η εκμάθηση συνάρτησης, η οποία αντιστοιχίζει τα χαρακτηριστικά της εισόδου στη σωστή ετικέτα εξόδου. Η επίδοση κρίνεται από την ικανότητα ορθής αντιστοίχισης μέσω της συνάρτησης που μαθαίνεται σε δεδομένα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση. Έτσι, κατά την εκπαίδευση οι αλγόριθμοι επιβλεπόμενης μάθησης στοχεύουν στη μείωση του σφάλματος στο σύνολο εκπαίδευσης, όμως η αξιολόγηση του μοντέλου κρίνεται με τον υπολογισμό των μετρικών σε διαφορετικό σύνολο δεδομένων, το σύνολο ελέγχου (test set) [14].

Η επιβλεπόμενη μάθηση έχει ευρεία εφαρμογή σε προβλήματα εύρεσης της κλάσης, από ένα γνωστό σύνολο κλάσεων, που ανήκει ένα δείγμα βάσει των χαρακτηριστικών του. Αυτή η κατηγορία είναι γνωστή ως ταξινόμηση (classification). Ειδική περίπτωση με ευρεία χρήση είναι η περίπτωση όπου τα δεδομένα είτε ανήκουν είτε δεν ανήκουν σε μία κλάση και είναι γνωστή ως δυϊκή ταξινόμηση (binary classification). Άλλη γενική κατηγορία επιβλεπόμενης μάθησης είναι αυτή, όπου ο στόχος είναι η πρόβλεψη μίας αριθμητικής τιμής, εφόσον έχει πραγματοποιηθεί η εκπαίδευση από ένα σύνολο γνωστών δεδομένων. Τέτοια προβλήματα είναι γνωστά ως προβλήματα παλινδρόμησης (regression).

Μη Επιβλεπόμενη Μάθηση

Οι αλγόριθμοι μη επιβλεπόμενης μάθησης δέχονται ως είσοδο ένα σύνολο δεδομένων, για τα οποία δεν δίνεται κάποιος χαρακτηρισμός, και στόχος είναι η εύρεση μοτίβων μεταξύ των δεδομένων, με σκοπό την εύρεση νέων αναπαραστάσεων ή την εκμάθηση παραγωγικών μοντέλων [14]. Χαρακτηριστική οικογένεια αλγορίθμων είναι αυτή της ανάλυσης κύριων συνιστωσών (Principal Component Analysis - PCA)[15], οι οποίοι στοχεύουν στη μείωση της διαστατικότητας της αναπαράστασης.

Η πιο γνωστή κατηγορία μη επιβλεπόμενης μάθησης είναι αυτή της ομαδοποίησης (clustering). Η ομαδοποίηση αποτελεί το πρόβλημα εύρεσης μιας διαμέρισης του αρχικού συνόλου, για την οποία τα παρόμοια δείγματα ανήκουν στο ίδιο σύνολο, ενώ τα ανόμοια αντικείμενα σε διαφορετικά [16]. Μία ακόμη σημαντική κατηγορία αποτελούν και οι αλγόριθμοι εύρεσης κανόνων συσχέτισης. Οι παραπάνω μέθοδοι έχουν ευρεία εφαρμογή σε προβλήματα ανίχνευσης ανωμαλιών [17], σε συστήματα προτάσεων (recommendation systems)[18] και σε άλλους τομείς, όπως η βιολογία [19].

Ενισχυτική Μάθηση

Στην ενισχυτική μάθηση η εκπαίδευση βασίζεται στις ανταμοιβές. Συγκεκριμένα, στόχος είναι να γίνει γνωστή η δράση, η οποία μεγιστοποιεί την ανταμοιβή. Για τον σκοπό αυτό ο πράκτορας (ή οι πράκτορες) αλληλεπιδρά με το περιβάλλον, ώστε να αποκτήσει πληροφορία για αυτό μέσω της ανταμοιβής που λαμβάνει. Αυτό που ξεχωρίζει την ενισχυτική μάθηση και την καθιστά ιδιαίτερα ενδιαφέρον τομέα είναι ότι λειτουργεί και σε περιπτώσεις όπου οι δράσεις δεν επηρεάζουν μόνο την άμεση ανταμοιβή, αλλά ολόκληρη την μελλοντική ακολουθία [20].

Η ενισχυτική μάθηση παρουσιάζει σημαντικές ομοιότητες με τη θεωρία ελέγχου, με αποτέλεσμα να παρέχει λύσεις βασισμένες σε δεδομένα σε πολλές εφαρμογές του πεδίου [21]. Τα δύο σημαντικότερα πεδία, όπου η ενισχυτική μάθηση βρίσκει ευρεία εφαρμογή σημειώνοντας αποτελέσματα σημαντικά καλύτερα από τους παραδοσιακούς αλγορίθμους είναι η ρομποτική [22] και τα ηλεκτρονικά παιχνίδια [23].

3.2 Νευρωνικά Δίκτυα

Η μάθηση στον ανθρώπινο οργανισμό λαμβάνει χώρα στα κύτταρα του νευρικού συστήματος, που είναι γνωστά ως νευρώνες (neurons). Τα εξωτερικά ερεθίσματα πραγματοποιούν αλλαγές στις συνάψεις (synapses) μεταξύ των νευρώνων και με αυτόν τον τρόπο επιτυγχάνεται η εκμάθηση στους ζωντανούς οργανισμούς. Την παραπάνω λειτουργία προσομοιώνουν τα τεχνητά νευρωνικά δίκτυα (artificial neural networks), με σκοπό να πραγματοποιηθεί η διαδικασία της μάθησης σε ένα υπολογιστικό σύστημα [24].

3.2.1 Δομή Νευρωνικών Δικτύων

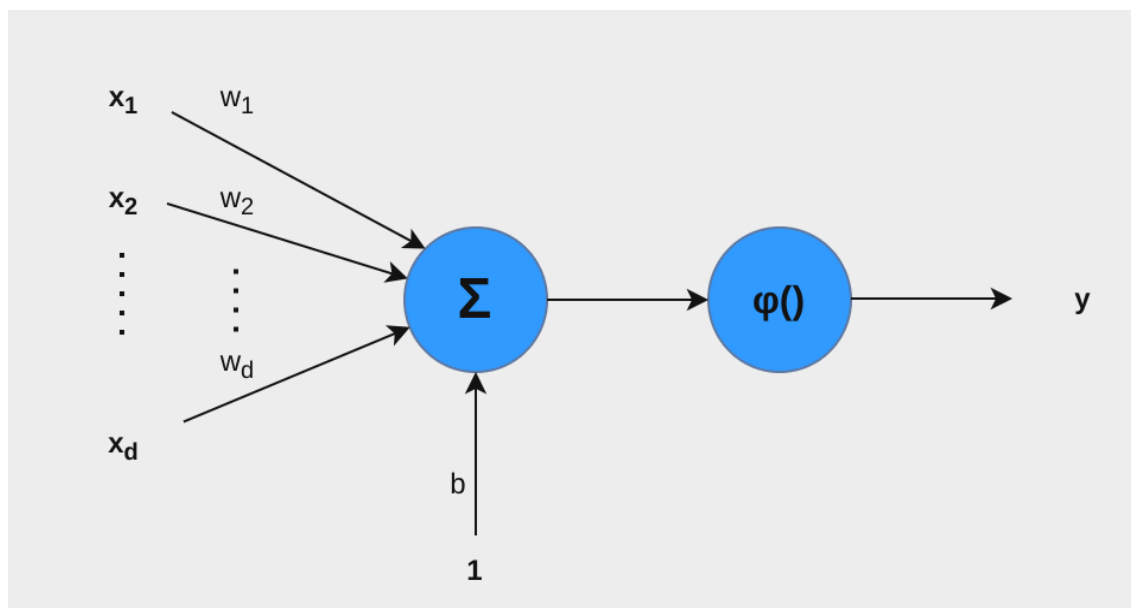
Οι βιολογικοί νευρώνες οργανώνονται σύμφωνα με τις συνδέσεις που πραγματοποιούνται μεταξύ των αξόνων και των δενδριτών. Οι συνδέσεις αυτές είναι γνωστές ως συνάψεις και αποτελούν τον θεμέλιο λίθο για τη διαδικασία της μάθησης στους έμβιους οργανισμούς. Συγκεκριμένα, στις αλλαγές των δυνάμεων των συνάψεων που παρατηρούνται λόγω εξωτερικών ερεθισμάτων οφείλεται η μάθηση.

Τα τεχνητά νευρωνικά δίκτυα έχουν ως στόχο την προσομοίωση της παραπάνω λειτουργίας για έναν υπολογιστή. Οι δυνάμεις που αντιστοιχούν στις συνάψεις αντιστοιχίζονται στα βάρη τα οποία πολλαπλασιάζουν την είσοδο, ενώ αντί των εξωτερικών ερεθισμάτων χρησιμοποιούνται τα δεδομένα εκπαίδευσης. Τέλος, χρησιμοποιείται μία συνάρτηση ενεργοποίησης, από την οποία υπολογίζεται η τελική έξοδος.

Πιο συγκεκριμένα, αν με x συμβολίζεται το διάνυσμα εισόδου, w το διάνυσμα των βαρών (το οποίο έχει τις ίδιες διαστάσεις με το διάνυσμα εισόδου), b ένας πραγματικός αριθμός και ϕ η συνάρτηση ενεργοποίησης, τότε το απλούστερο μοντέλο νευρωνικού δικτύου, το οποίο είναι γνωστό και ως νευρώνας perceptron, υπολογίζει τη συνάρτηση:

$$y = \phi(w^T \cdot x + b)$$

Είναι εμφανές ότι το όρισμα της συνάρτησης ενεργοποίησης αποτελεί αφινική συνάρ-



Σχήμα 3.1: Τεχνητό Νευρωνικό Δίκτυο

τηση της εισόδου, με αποτέλεσμα ένα τέτοιο μοντέλο να είναι περιορισμένο, αποτελώντας καλή περίπτωση για προβλήματα με γραμμικά διαχωρίσιμα δεδομένα. Για την επίλυση του προβλήματος αυτό, τα μοντέλα νευρωνικών δικτύων οργανώνονται σε πολυπλοκότερες αρχιτεκτονικές, ώστε να διευρυθούν οι υπολογιστικές τους δυνατότητες.

3.2.2 Αλγόριθμος Perceptron

Η παραπάνω απλή εκδοχή ενός νευρωνικού δικτύου μπορεί να εκπαιδευθεί με χρήση του αλγορίθμου perceptron. Πρόκειται για έναν επαναληπτικό αλγόριθμο, ο οποίος σε κάθε επανάληψη προσπαθεί να μειώσει τον αριθμό των σφαλμένων προβλέψεων στο σύνολο εκπαίδευσης. Στην περίπτωση που τα δεδομένα είναι γραμμικά διαχωρίσιμα αποδεικνύεται θεωρητικά η σύγκλιση του αλγορίθμου. Ακολουθεί η περιγραφή του αλγορίθμου.

Αρχικά, τα βάρη αρχικοποιούνται σε κάποιες τιμές, για παράδειγμα όλα τα βάρη μπορούν να θεωρηθούν μηδενικά. Στην t -οστή επανάληψη, τα βάρη ανανεώνονται με την τιμή $w^{(t)} \leftarrow w^{(t-1)} + \eta y_i \cdot x_i$, για κάθε δείγμα i , όπου ισχύει $y_i \cdot (w^{(t)T} \cdot x_i + b) < 0$, δηλαδή η πρόβλεψη είναι λανθασμένη. Ο αριθμός η , ο οποίος ονομάζεται ρυθμός μάθησης (learning rate), είναι θετικός και από την τιμή του εξαρτάται το πλήθος βημάτων που απαιτούνται για τη σύγκλιση. Εφόσον τα δεδομένα είναι γραμμικά διαχωρίσιμα, τότε ο αλγόριθμος τερματίζει όταν δεν υπάρχει κανένα σφάλμα σε μία επανάληψη. Διαφορετικά, ορίζεται μέγιστο πλήθος επαναλήψεων, ώστε να διασφαλιστεί ο τερματισμός[16].

3.2.3 Αρχιτεκτονικές Νευρωνικών Δικτύων

Το μοντέλο του νευρώνα perceptron είναι ιδιαίτερα απλό και κατ' επέκταση οι δυνατότητες του περιορισμένες, ώστε να λειτουργεί κυρίως για γραμμικές συναρτήσεις. Τα δίκτυα πρόσθιας τροφοδότησης (feedforward neural networks), γνωστά και ως πολυεπίπεδο Perceptron (multilayer perceptron - MLP) συνδυάζουν νευρώνες perceptron σε πολλαπλά επίπεδα,

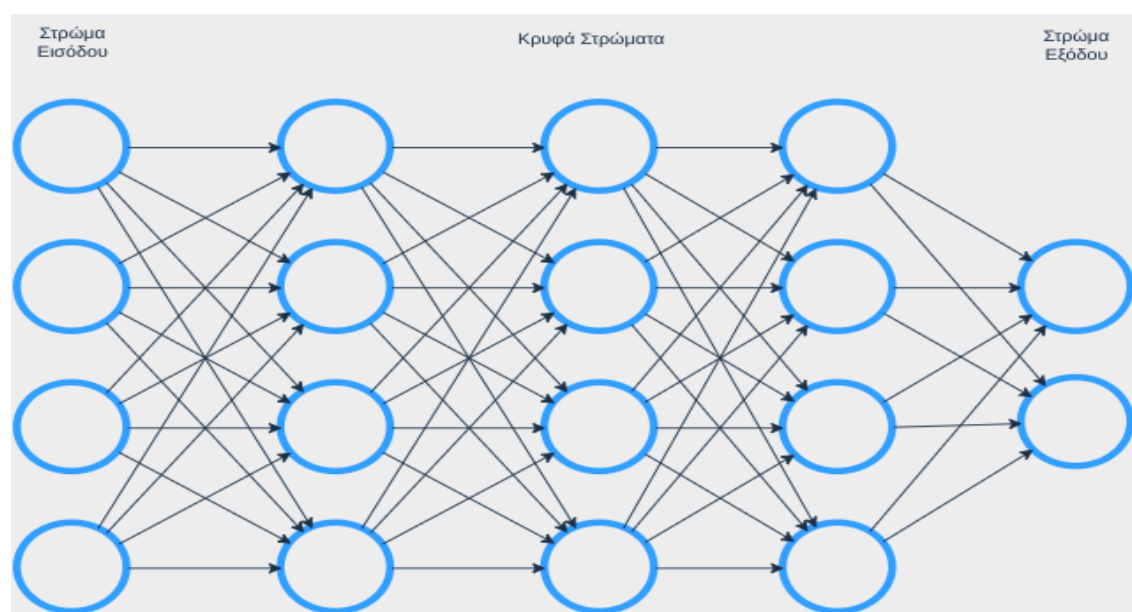
ώστε να αυξηθούν οι υπολογιστικές δυνατότητες. Τα δίκτυα αυτά, αποτελούν το κύριο εργαλείο της βαθιάς μάθησης, η οποία παρουσιάζεται εκτενέστερα στην επόμενη ενότητα. Μία ειδική κατηγορία των MLPs αποτελούν τα συνελκτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNN), τα οποία χρησιμοποιούν λιγότερες συνδέσεις μεταξύ των επιπέδων, προσομοιώνοντας την πράξη της συνέλιξης. Τα CNNs έχουν ευρεία εφαρμογή σε προβλήματα αναγνώρισης αντικειμένων σε εικόνες. Τέλος, ιδιαίτερα σημαντική κατηγορία νευρωνικών δικτύων αποτελούν τα επαναληπτικά νευρωνικά δίκτυα (Recurrent Neural Networks - RNN), τα οποία επιτρέπουν την ανατροφοδότηση της εξόδου στην είσοδο του ίδιου νευρωνικού και τα οποία έχουν εκτεταμένες χρήσεις στην επεξεργασία φυσικής γλώσσας [25].

3.3 Βαθιά Μάθηση

Η βαθιά μάθηση (deep learning) αποτελεί κατηγορία της μηχανικής μάθησης, η οποία χρησιμοποιεί δίκτυα πρόσθιας τροφοδότησης πολλαπλών επιπέδων, προκειμένου να μάθει μία συνάρτηση που αντιστοιχεί τα δεδομένα εισόδου στην επιθυμητή έξοδο.

3.3.1 Δίκτυα πρόσθιας τροφοδότησης

Τα δίκτυα πρόσθιας τροφοδότησης (feedforward networks) ή MLP αποτελούν το θεμέλιο του πεδίου της βαθιάς μάθησης. Αποτελούνται από νευρώνες perceptron, οι οποίοι οργανώνονται σε επίπεδα, όπου η έξοδος κάθε επιπέδου δίνεται ως είσοδος σε κάθε νευρώνα του αμέσως επόμενου επιπέδου. Το τελικό επίπεδο ονομάζεται στρώμα εξόδου (output layer) και καθορίζουν τη μορφή της εξόδου, ενώ τα ενδιάμεσα επίπεδα αποτελούν τα κρυφά στρώματα (hidden layers). Μέσω αυτής της οργάνωσης αυξάνονται σημαντικά οι δυνατότητες για την προσέγγιση συναρτήσεων. Στη βαθιά μάθηση χρησιμοποιούνται τυπικά τουλάχιστον τρία επίπεδα, καθώς αυξάνοντας το βάθος του δικτύου είναι πιθανότερο να αρκούν λιγότερες μονάδες perceptron για την αναπαράσταση της επιθυμητής συνάρτησης.



Σχήμα 3.2: Αρχιτεκτονική Δικτύου Perceptron Πολλαπλών Επιπέδων

3.3.2 Συνάρτηση ενεργοποίησης

Η συνάρτηση ενεργοποίησης που χρησιμοποιείται σε κάθε υπολογιστική μονάδα perceptron αποτελεί θεμελιώδη κεφάλαιο για τη βαθιά μάθηση. Μέσω της συνάρτησης αυτής εισάγεται μη γραμμικότητα κατά τον υπολογισμό της εξόδου και αυτός είναι ο λόγος που τα MLP επιτυγχάνουν να προσεγγίσουν μη γραμμικές συναρτήσεις. Αποδεικνύεται, μάλιστα, ότι ένα νευρωνικό δίκτυο το οποίο αποτελείται από δύο μόλις επίπεδα μπορεί να προσεγγίσει οποιαδήποτε συνεχή και φραγμένη συνάρτηση κατά οποιοδήποτε θετικό σφάλμα, όμως ο αριθμός των μονάδων σε κάθε επίπεδο μπορεί να χρειάζεται να είναι εκθετικά μεγάλος.

Μία συνάρτηση ενεργοποίησης, η οποία χρησιμοποιείται ευρέως, είναι η Rectified Linear Unit (ReLU), η οποία ορίζεται ως

$$ReLU(z) = \max\{0, z\} \quad (3.1)$$

Ουσιαστικά, η συνάρτηση αυτή είναι ίδια με την ταυτοτική στην ενεργό περιοχή της και μηδέν αλλού. Το κύριο πλεονέκτημα της συνάρτησης είναι ότι η παράγωγος της είναι 1 στην ενεργή περιοχή και 0 σε όλα τα άλλα σημεία που ορίζεται, ενώ η δεύτερη παράγωγος είναι μηδέν όπου ορίζεται. Αυτό διευκολύνει σημαντικά τη διαδικασία της εκπαίδευσης μέσω των αλγορίθμων βελτιστοποίησης που παρουσιάζονται στη συνέχεια, καθιστώντας την συνάρτηση αυτή ως την καλύτερη επιλογή για τα ενδιάμεσα στρώματα.

Εναλλακτικές συναρτήσεις ενεργοποίησης είναι η λογιστική σιγμοειδής (logistic sigmoid) και η υπερβολική εφαπτομένη (hyperbolic tangent), οι οποίες ορίζονται ως

$$\sigma(z) = \frac{1}{1 + e^{-x}} \quad (3.2)$$

$$\tanh(z) = 2\sigma(2z) - 1 \quad (3.3)$$

Οι συναρτήσεις αυτές έχουν το μειονέκτημα ότι η παράγωγος είναι περίπου μηδέν στο μεγαλύτερο μέρος του πεδίου που ορίζονται. Γι' αυτό το λόγο η χρήση τους σε ενδιάμεσες μονάδες δυσχεραίνει την εκπαίδευση, οπότε συνήθως αποφεύγεται. Όμως χρησιμοποιούνται ευρέως ως συναρτήσεις ενεργοποίησης στις μονάδες του στρώματος εξόδου.

3.3.3 Εκπαίδευση MLP

Σχετικά με την εκπαίδευση πρέπει αρχικά να οριστεί η συνάρτηση κόστους (cost function), η οποία πρόκειται να ελαχιστοποιηθεί. Συνηθισμένες οικογένειες συναρτήσεων που χρησιμοποιούνται ως κόστη είναι οι συναρτήσεις μέσου τετραγωνικού σφάλματος (MSE - Mean Squared Error) και συναρτήσεις απόστασης μεταξύ κατανομών (για παράδειγμα, cross entropy). Η συνάρτηση αυτή εκφράζει την απόσταση των εκτιμήσεων του μοντέλου που εκπαιδεύεται εκείνη τη στιγμή σε σχέση με τα πραγματικά δεδομένα.

Η εκπαίδευση βασίζεται στον αλγόριθμο Stochastic Gradient Descent. Σύμφωνα με τους αλγορίθμους της οικογένειας Gradient Descent η τιμή των παραμέτρων που βελτιστοποιούνται, στην προκειμένη περίπτωση τα βάρη των νευρωνικών δικτύων, ανανεώνεται προς την κατεύθυνση που μειώνεται το κόστος. Η κατεύθυνση αυτή είναι αντίθετη από την κλίση (gradient) της συνάρτησης κόστους στο σημείο που βρίσκοταν εκείνη τη χρονική στιγμή,

δηλαδή αν w οι παράμετροι προς εκμάθηση και J η συνάρτηση κόστους, τότε η ανανέωση γίνεται βάσει της σχέσης:

$$w \leftarrow w - \eta \nabla_w J(w) \quad (3.4)$$

Όπου το η ονομάζεται ρυθμός μάθησης (learning rate) και είναι σημαντική παράμετρος της εκπαίδευσης. Μικρή τιμή του έχει κίνδυνο να απαιτούνται πολλές επαναλήψεις, ώστε να βρεθεί η βέλτιστη λύση, επιβραδύνοντας έτσι τη διαδικασία της εκπαίδευσης. Αντίθετα, μεγάλη τιμή του έχει ως κίνδυνο να γίνονται ταλαντώσεις και να μην προσεγγιστεί ικανοποιητικά η βέλτιστη λύση. Επειδή το κόστος εξαρτάται από τη συνάρτηση που έχει μάθει σε εκείνο το βήμα το δίκτυο, ο υπολογισμός της κλίσης δεν είναι εύκολος. Γι' αυτόν τον λόγο υπολογίζεται μία προσέγγιση της κλίσης σύμφωνα με τα δεδομένα εκπαίδευσης. Στον αλγόριθμο Stochastic Gradient Descent επιλέγεται ένα μικρό σύνολο από δεδομένα για τον υπολογισμό της κλίσης.

Λόγω της πολύπλοκης αρχιτεκτονικής που προκύπτει σε δίκτυα με πολλαπλά επίπεδα και πιθανώς πολλές μονάδες σε κάθε επίπεδο, για τον υπολογισμό της κλίσης χρησιμοποιείται ο αλγόριθμος backpropagation. Πρόκειται για έναν γενικό αλγόριθμο υπολογισμού της κλίσης συναρτήσεων, ο οποίος στηρίζεται στον κανόνα της αλυσίδας. Μέσω του αλγορίθμου αποφεύγονται επαναλήψεις υπολογισμών, με αποτέλεσμα να γίνονται αποδοτικά οι ζητούμενοι υπολογισμοί.

3.4 Μαρκοβιανές Διαδικασίες Απόφασης

Στην παρούσα ενότητα παρουσιάζονται οι έννοιες των Μαρκοβιανών Διαδικασιών Απόφασης (ΜΔΑ) (Markov Decision Processes - MDP) και των Μαρκοβιανών Παιγνίων (Markov Games), οι οποίες αποτελούν κατάλληλα πλαίσια για τη μοντελοποίηση των προβλημάτων ενισχυτικής μάθησης.

3.4.1 Θεωρία Μαρκοβιανών Διαδικασιών Απόφασης

Οι Μαρκοβιανές Διαδικασίες Απόφασης αποτελούν συνήθη μέθοδο τυποποίησης προβλημάτων σειριακής λήψης αποφάσεων, για περιπτώσεις όπου η τρέχουσα επιλογή επηρεάζει και τις μελλοντικές ανταμοιβές. Έτσι, αποτελούν εξιδανικευμένη περίπτωση της ενισχυτικής μάθησης, όπου μπορούν να υπάρξουν θεωρητικές τεκμηριώσεις[20].

Στο πλαίσιο αυτό, ως πράκτορας (agent) θεωρείται το υπολογιστικό σύστημα που εκπαιδεύεται σχετικά με τη λήψη αποφάσεων. Οτιδήποτε δεν μπορεί να τροποποιηθεί από τον πράκτορα θεωρείται ότι ανήκει στο περιβάλλον (environment). Το περιβάλλον δηλώνει την αδυναμία ελέγχου από πλευράς του πράκτορα χωρίς όμως να δηλώνει την έλλειψη γνώσης γι' αυτό. Σε πολλές περιπτώσεις ο πράκτορας έχει πληροφορία για την κατάσταση του ή μπορεί ακόμη και ένα μέρος του περιβάλλοντος να ανήκει στο ίδιο φυσικό σύστημα με τον πράκτορα [20].

Οι Μαρκοβιανές Διαδικασίες Απόφασης ορίζονται ως μία πλειάδα $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, όπου:

- $\mathcal{S}$ είναι το σύνολο των καταστάσεων

- $\mathcal{A}$ είναι το σύνολο των δράσεων (αποφάσεων)
- $\mathcal{P}$ είναι η συνάρτηση που περιγράφει την κατανομή της πιθανότητας μετάβασης από μία κατάσταση s_t , όταν εκτελεστεί η δράση a_t
- $\mathcal{R}$ η συνάρτηση ανταμοιβής
- γ ο παράγοντας μείωσης

Σημειώνεται ακόμη ότι κάθε κατάσταση που ανήκει στο σύνολο $\mathcal{S}$ ικανοποιεί τη μαρκοβιανή ιδιότητα, δηλαδή αποτελεί πλήρη περιγραφή της τρέχουσας κατάστασης και οι παρελθούσες καταστάσεις δεν έχουν επιρροή[26].

Σχετικά με τη συνάρτηση $\mathcal{P}$, αυτή ορίζεται ως:

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

και δηλώνει την πιθανότητα μετάβασης στην κατάσταση s' την επόμενη χρονική στιγμή, εφόσον ο πράκτορας βρίσκεται στην κατάσταση s και εκτελεστεί η δράση a . Πρόκειται για μία ντερμινιστική συνάρτηση, η οποία καθορίζει τη δυναμική του προβλήματος ΜΔΑ[20].

Ο αριθμός γ ονομάζεται παράγοντας μείωσης (discount factor) και είναι ένας πραγματικός αριθμός που ανήκει στο διάστημα $[0, 1]$. Η τιμή 0 δηλώνει ότι οι αποφάσεις του πράκτορα οφείλουν να λαμβάνουν υπόψιν μόνο την ανταμοιβή που θα ληφθεί την τρέχουσα χρονική στιγμή, ενώ αυξάνοντας την τιμή του αυξάνεται αντίστοιχα το βάρος των μελλοντικών ανταμοιβών στη λήψη αποφάσεων.

Η συνάρτηση ανταμοιβής $\mathcal{R}$ υπολογίζει την αναμενόμενη ανταμοιβή, όταν σε μία κατάσταση s εκτελεστεί η δράση a . Συγκεκριμένα, είναι $R_s^a = \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a]$. Ωστόσο, ο στόχος της ενισχυτικής μάθησης είναι να μεγιστοποιήσει τη συνολική ανταμοιβή που θα λάβει ο πράκτορας σε ένα επεισόδιο. Γί αυτό τον λόγο εισάγεται η συνάρτηση αναμενόμενης ανταμοιβής (expected return) $\mathcal{G}_t$. Θεωρώντας ότι η ακολουθία αποφάσεων ολοκληρώνεται στο βήμα T , για παράγοντα μείωσης γ , η συνάρτηση αυτή ορίζεται ως $\mathcal{G}_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T = R_{t+1} + \gamma \mathcal{G}_{t+1}$. Μία τέτοια ακολουθία αποφάσεων, η οποία ολοκληρώνεται μετά από T βήματα, είναι συχνά παρατηρούμενη σε προβλήματα ενισχυτικής μάθησης και αποτελεί ένα επεισόδιο. Υπάρχουν, όμως, και περιπτώσεις όπου δεν υπάρχει τελική κατάσταση και μπορεί η ακολουθία των αποφάσεων να είναι άπειρη. Σε τέτοιες περιπτώσεις, η συνάρτηση $\mathcal{G}_t$ ορίζεται ανάλογα ως $\mathcal{G}_t = \sum_{i=1}^{+\infty} \gamma^{i-1} R_{t+i} = R_{t+1} + \gamma \mathcal{G}_{t+1}$, για $\gamma \in (0, 1)$. Για $\gamma = 0$ ισχύει $\mathcal{G}_t = R_{t+1}$, ενώ για $\gamma = 1$ μπορεί να χρησιμοποιηθεί ο παραπάνω ορισμός, όμως υπάρχει ο κίνδυνος ακόμη και για πεπερασμένες αμοιβές η συνάρτηση να απειρίζεται [20].

3.4.2 Μαρκοβιανά Παίγνια

Τα Μαρκοβιανά Παίγνια αποτελούν μια γενικότερη προσέγγιση από τις Μαρκοβιανές Διαδικασίες Απόφασης και είναι κατάλληλα πλαίσια περιγραφής προβλημάτων ενισχυτικής μάθησης πολλαπλών πρακτόρων. Για τον ορισμό ενός μαρκοβιανού παιγνίου χρησιμοποιούνται οι εξής έννοιες:

- k το πλήθος των πρακτόρων
- $\mathcal{S}$, το σύνολο των καταστάσεων
- $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$, το σύνολο των δράσεων του κάθε πράκτορα
- $\mathcal{T}$, η συνάρτηση η μετάβασης
- $\mathcal{R}_i$, η συνάρτηση ανταμοιβής του πράκτορα i

Αν οριστεί ως $PD(\mathcal{S})$ το σύνολο των κατανομών πιθανότητας πάνω στο σύνολο $\mathcal{S}$, τότε η συνάρτηση μετάβασης $\mathcal{T}$ ορίζεται ως $\mathcal{T} : \mathcal{S} \times \mathcal{A}_1 \times \dots \times \mathcal{A}_k \rightarrow PD(\mathcal{S})$. Αντίστοιχα ορίζεται για κάθε πράκτορα η συνάρτηση ανταμοιβής $\mathcal{R}_i : \mathcal{S} \times \mathcal{A}_1 \times \dots \times \mathcal{A}_k \rightarrow \mathbb{R}$. Ο στόχος κάθε πράκτορα είναι η μεγιστοποίηση της αναμενόμενης συνολικής ανταμοιβής $\mathbb{E}\{\sum_{j=0}^{+\infty} \gamma^j r_{i,t+j}\}$, όπου $r_{i,t+k}$ η ανταμοιβή που λαμβάνει ο πράκτορας i μετά από j βήματα και γ ο παράγοντας μείωσης, όπως ορίστηκε για τις ΜΔΑ [27].

3.5 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση αποτελεί μία από τις κατηγορίες μηχανικής μάθησης, η οποία έχει ως στόχο ένας πράκτορας να μάθει μία πολιτική, η οποία μεγιστοποιεί την ανταμοιβή που θα λάβει. Στην παρούσα εργασία θεωρείται ότι το πρόβλημα της ενισχυτικής μάθησης περιγράφεται από μία Μαρκοβιανή Διαδικασία Απόφασης, όπως αυτή ορίστηκε στην προηγούμενη ενότητα. Σημειώνεται ότι οι ΜΔΑ δεν είναι ο μοναδικός τρόπος αναπαράστασης των προβλημάτων που εφαρμόζονται τεχνικές ενισχυτικής μάθησης.

3.5.1 Βασικοί Ορισμοί

Η πολιτική (policy) ορίζεται ως η κατανομή πιθανότητας να επιλεγεί μία δράση a από το σύνολο $\mathcal{A}$, εφόσον ο πράκτορας βρίσκεται στην κατάσταση s του συνόλου $\mathcal{S}$. Η πολιτική συμβολίζεται με το γράμμα π και με $\pi(a|s)$ δηλώνεται η πιθανότητα να επιλεγεί η δράση a , αν τη δεδομένη χρονική στιγμή t ο πράκτορας βρίσκεται στην κατάσταση s και ακολουθεί την πολιτική π . Σκοπός της ενισχυτικής μάθησης είναι να βρεθεί η πολιτική που μεγιστοποιεί την αναμενόμενη ανταμοιβή, μέσω της εμπειρίας που λαμβάνεται από την αλληλεπίδραση με το περιβάλλον [20].

Για κάθε κατάσταση $s \in \mathcal{S}$ και μία πολιτική π ορίζεται η συνάρτηση αξίας-κατάστασης (state-value function), η οποία δηλώνει την αναμενόμενη ανταμοιβή, όταν ο πράκτορας ξεκινάει από την κατάσταση s και ακολουθεί την πολιτική π . Για τις τελικές καταστάσεις, εφόσον υπάρχουν, η τιμή της συνάρτησης είναι μηδέν, ενώ για τις υπόλοιπες ο υπολογισμός γίνεται σύμφωνα με τον τύπο:

$$v_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{+\infty} \gamma^k R_{t+k+1} | S_t = s \right], \forall s \in \mathcal{S}$$

Αντίστοιχη συνάρτηση μπορεί να οριστεί για να δηλώσει την αναμενόμενη ανταμοιβή, όταν ο πράκτορας βρίσκεται στην κατάσταση s και εκτελεί τη δράση a . Σε αυτή την περίπτωση

γίνεται λόγος για συνάρτηση αξίας-δράσης (action-value function). Για τη συνάρτηση αυτή, ισχύει η εξίσωση:

$$q_{\pi}(s, a) = \mathbb{E}_{\pi} \left[\gamma^k R_{t+k+1} | S_t = s, A_t = a \right]$$

Οι περισσότεροι αλγόριθμοι ενισχυτικής μάθησης έχουν ως στόχο τον υπολογισμό της συνάρτησης αξίας-κατάστασης ή τη συνάρτηση αξίας-δράσης μέσω των δεδομένων που λαμβάνουν από το περιβάλλον. Για τον υπολογισμό χρησιμοποιούνται δύο κύριοι μέθοδοι: ο δυναμικός προγραμματισμός και η μέθοδος Monte-Carlo.

3.5.2 Κατηγορίες Ενισχυτικής Μάθησης

Παρακάτω παρουσιάζονται τα βασικά κριτήρια βάσει των οποίων ταξινομούνται οι αλγόριθμοι ενισχυτικής μάθησης. Σημειώνεται ότι τα ζεύγη κατηγοριών που παρουσιάζονται αφορούν ένα συγκεκριμένο χαρακτηριστικό του αλγορίθμου, οπότε ένας αλγόριθμος θα ανήκει το πολύ σε μία από τις δύο κατηγορίες του ζεύγους, όμως μπορεί να ανήκει και σε περισσότερες κατηγορίες που αναλύονται παρακάτω.

Model-free / Model-based

Πλήθος αλγορίθμων χρησιμοποιεί ένα μοντέλο για την περιγραφή του περιβάλλοντος. Τέτοιοι αλγόριθμοι, όπως ο AlphaZero, μπορούν να εκμεταλλευτούν το μοντέλο για να έχουν περισσότερη πληροφορία για τις μελλοντικές καταστάσεις και να λάβουν υπόψιν αυτή την πληροφορία στην πολιτική που μαθαίνουν. Το κυριότερο μειονέκτημα αυτής της μεθόδου είναι ότι δεν υπάρχει πάντα αρκετά ακριβές μοντέλο που να προσομοιώνει το περιβάλλον, με αποτέλεσμα να λαμβάνονται μη βέλτιστα αποτελέσματα. Από την άλλη, η οικογένεια των model-free αλγορίθμων δεν προσπαθεί να μάθει τη δυναμική του περιβάλλοντος άμεσα, με αποτέλεσμα να είναι απλούστερη η υλοποίηση τους και κατ' επέκταση ευρύτερη η χρήση τους [28].

Online / Offline

Στην ενισχυτική μάθηση τα δεδομένα προκύπτουν μέσω της αλληλεπίδρασης με το περιβάλλον. Η συνηθέστερη περίπτωση είναι η αλληλεπίδραση να γίνεται παράλληλα με τη διαδικασία της εκπαίδευσης, οπότε τυχόν αλλαγές στο περιβάλλον να γίνονται άμεσα αντιληπτές. Αυτή αποτελεί την οικογένεια των online αλγορίθμων. Αντίθετα, στους offline αλγορίθμους ενισχυτικής μάθησης τα δεδομένα είναι υπολογισμένα εκ των προτέρων. Αυτό έχει ως αποτέλεσμα αυξημένη αξιοπιστία και ταχύτερη σύγκλιση, όμως η ακρίβεια εξαρτάται από τα αρχικά δεδομένα, οπότε δεν είναι κατάλληλη για περιβάλλοντα που μεταβάλλονται γρήγορα.

On-policy / Off-policy

Στην online ενισχυτική μάθηση η εκπαίδευση και η αλληλεπίδραση γίνονται παράλληλα. Μία συνηθισμένη πρακτική είναι στην αλληλεπίδραση με το περιβάλλον να χρησιμοποιείται η πολιτική που έχει μάθει ο πράκτορας κατά την εκπαίδευση εκείνη τη χρονική στιγμή. Αυτή η πρακτική αποτελεί την on-policy ενισχυτική μάθηση. Από την άλλη, οι off-policy αλγόριθμοι χρησιμοποιούν διαφορετική πολιτική για την αλληλεπίδραση. Αυτή η μέθοδος μπορεί να εκμεταλλευτεί τη διαφορετική πολιτική, για να βελτιώσει την εξερεύνηση προς την αναζήτηση της βέλτιστης λύσης.

3.5.3 Εκμάθηση Συναρτήσεων

Η κεντρική ιδέα της ενισχυτικής μάθησης είναι η εκμάθηση της συνάρτησης αξίας-κατάστασης (state-value function) και της συνάρτησης αξίας-δράσης (action-value function) και η εύρεση της βέλτιστης πολιτικής μέσω αυτών. Σχετικά με τις βέλτιστες συναρτήσεις αξίας είναι γνωστό ότι ικανοποιούν τις εξισώσεις βελτιστότητας του Bellman [20]:

$$\begin{aligned} v_*(s) &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\ &= \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v_*(s')] \end{aligned} \quad (3.5)$$

$$\begin{aligned} q_*(s, a) &= \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a] \\ &= \sum_{s', r} p(s', r | s, a) [r + \gamma \max_{a'} q_*(s', a')] \end{aligned} \quad (3.6)$$

Δυναμικός Προγραμματισμός

Ο δυναμικός προγραμματισμός αποτελεί μία γενική μέθοδο σχεδιασμού αλγορίθμων, η οποία στηρίζεται στη χρήση μνήμης για την αποθήκευση των βέλτιστων λύσεων για ενδιαμέσα υποπροβλήματα. Η εκμάθηση των συναρτήσεων αξίας μπορεί να πραγματοποιηθεί μέσω του δυναμικού προγραμματισμού, εφόσον είναι γνωστή η δυναμική του περιβάλλοντος. Αυτό καθιστά δύσκολη την πρακτική εφαρμογή των συγκεκριμένων αλγορίθμων, όμως στην ιδέα αυτή βασίζονται οι πιο διαδεδομένοι αλγόριθμοι ενισχυτικής μάθησης.

Σύμφωνα με τον δυναμικό προγραμματισμό, η συνάρτηση αξίας-κατάστασης για μία συγκεκριμένη πολιτική π μπορεί να υπολογιστεί με τον ακόλουθο αλγόριθμο. Αρχικά, επιλέγεται αυθαίρετα μία τιμή v_0 για κάθε κατάσταση, με μοναδικό περιορισμό η τιμή αυτή να είναι 0 για τελικές καταστάσεις. Στη συνέχεια, σε κάθε επανάληψη η τιμή αυτή ανανεώνεται σύμφωνα με τον κανόνα:

$$\begin{aligned} v_{k+1}(s) &\leftarrow \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s], \text{ ή} \\ v_{k+1}(s) &\leftarrow \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) [r + \gamma v_k(s')] \end{aligned} \quad (3.7)$$

Για κάθε κατάσταση s . Αποδεικνύεται ότι η ακολουθία $\{v_k\}$ συγκλίνει στη συνάρτηση αξίας-κατάστασης υπό την πολιτική π v_π . Αυτός ο αλγόριθμος ονομάζεται policy value iteration.

Ο παραπάνω αλγόριθμος μαθαίνει τη συνάρτηση αξίας-κατάστασης για μία συγκεκριμένη πολιτική. Όμως, ο στόχος της ενισχυτικής μάθησης είναι η εύρεση της βέλτιστης πολιτικής. Επομένως, ο παραπάνω αλγόριθμος δεν μπορεί να χρησιμοποιηθεί από μόνος του, καθώς η εύρεση της συνάρτησης αξίας-κατάστασης για κάθε δυνατή πολιτική με βάση τον προηγούμενο αλγόριθμο είναι πρακτικά ασύμφορη έως και αδύνατη, ακόμα και για λογικό πλήθος καταστάσεων και δράσεων.

Η λύση στο πρόβλημα αυτό εισέρχεται μέσω του θεωρήματος βελτίωσης πολιτικής (policy improvement theorem). Σύμφωνα με το θεώρημα αυτό, αν π και π' είναι δύο ντετερμινιστικές πολιτικές τέτοιες, ώστε για κάθε κατάσταση $s \in \mathcal{S}$ να ισχύει ότι,

$$q_\pi(s, \pi'(s)) \geq v_\pi(s), \quad (3.8)$$

τότε ισχύει ότι η πολιτική π' είναι τουλάχιστον τόσο καλή όσο η πολιτική π , δηλαδή για κάθε κατάσταση $s \in \mathcal{S}$ ισχύει ότι,

$$v_{\pi'}(s) = v_{\pi}(s). \quad (3.9)$$

Σημειώνεται ότι η τιμή $q_{\pi}(s, a)$ δηλώνει την αναμενόμενη ανταμοιβή αν ο πράκτορας βρίσκεται στην κατάσταση s , εκτελέσει τη δράση a και στη συνέχεια ακολουθεί την πολιτική π , και υπολογίζεται σύμφωνα με την εξίσωση:

$$q_{\pi}(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma v_{\pi}(s')] \quad (3.10)$$

Βάσει του παραπάνω θεωρήματος η επιλογή νέας πολιτικής, σύμφωνα με τη σχέση:

$$\pi'(s) = \operatorname{argmax}_a q_{\pi}(s, a) \quad (3.11)$$

αποτελεί μία άπληστη μέθοδος εύρεσης μίας καλύτερης πολιτικής.

Συνδυάζοντας τα παραπάνω προκύπτει ο αλγόριθμος policy iteration, ο οποίος χρησιμοποιεί μία αυθαίρετη αρχικοποίηση και σε κάθε βήμα εκτελεί αρχικά τον αλγόριθμο policy evaluation και στη συνέχεια τον αλγόριθμο policy improvement. Ο αλγόριθμος τερματίζει όταν ο αλγόριθμος policy improvement δεν μεταβάλλει την πολιτική σε κάποιο βήμα. Η συνθήκη τερματισμού διασφαλίζει ότι η πολιτική που προκύπτει ικανοποιεί την εξίσωση Bellman και επομένως είναι η βέλτιστη[20].

Ο παραπάνω αλγόριθμος έχει το μειονέκτημα ότι σε κάθε βήμα εκτελείται πλήρως ο αλγόριθμος policy evaluation, το οποίο μπορεί να οδηγήσει σε αυξημένο υπολογιστικό χρόνο. Μία βελτίωση είναι ο αλγόριθμος value iteration, ο οποίος χρησιμοποιεί σε κάθε βήμα ένα μόνο πέρασμα για κάθε κατάσταση από τον αλγόριθμο policy evaluation και εκτελεί άμεσα τον αλγόριθμο policy improvement. Συγκεκριμένα, για τις ίδιες υποθέσεις σχετικά με την αρχικοποίηση των συναρτήσεων, σε κάθε βήμα η συνάρτηση κατάστασης-αξίας ανανεώνεται σύμφωνα με τη σχέση:

$$v_{k+1} = \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v_k(s')] \quad (3.12)$$

Ως συνθήκη τερματισμού χρησιμοποιείται η τιμή της συνολικής μεταβολής $\sum_s |v_{k+1}(s) - v_k(s)|$ να είναι μικρότερη από μία επιλεγμένη τιμή και ως έξοδος θεωρείται η πολιτική:

$$\pi(s) = \operatorname{argmax}_a \sum_{s', r} p(s', r | s, a) [r + \gamma V(s')] \quad (3.13)$$

Όπου με $V(s)$ η τιμή της συνάρτησης αξίας-κατάστασης, όταν τερματίζει ο αλγόριθμος. Ο αλγόριθμος αυτός επίσης υπολογίζει τη βέλτιστη πολιτική, όμως συγκλίνει γρηγορότερα σε σχέση με τον policy iteration.

Μέθοδος Monte-Carlo

Οι μέθοδοι που βασίζονται στον δυναμικό προγραμματισμό έχουν τα μειονεκτήματα ότι πρέπει να είναι γνωστή η δυναμική του περιβάλλοντος και ότι είναι απαραίτητος ο υπολογισμός της αξίας για κάθε δυνατή κατάσταση. Αντίθετα, οι μέθοδοι Monte-Carlo μπορούν

να υπολογίσουν βέλτιστες πολιτικές χρησιμοποιώντας δείγματα που προκύπτουν είτε μέσω πραγματικής αλληλεπίδρασης με το περιβάλλον είτε μέσω προσομοιώσεων. Ακόμη, μπορούν να χρησιμοποιηθούν για τον υπολογισμό της συνάρτησης αξίας-κατάστασης για μία μόνο κατάσταση ή για ένα υποσύνολο καταστάσεων[20].

Ο όρος μέθοδος Monte-Carlo χρησιμοποιείται για να περιγράψει αλγόριθμους, οι οποίοι χρησιμοποιούν τυχειότητα για τον υπολογισμό του ζητούμενου. Έτσι, στην ενισχυτική μάθηση μπορεί να χρησιμοποιηθεί η μέθοδος Monte-Carlo για τον υπολογισμό της συνάρτησης αξίας-κατάστασης βάσει της μέσης αναμενόμενης ανταμοιβής, όπως υπολογίζεται από κάποια δείγματα. Τα δείγματα αυτά μπορούν να έχουν προκύψει από πραγματικές αλληλεπιδράσεις με το περιβάλλον ή από προσομοιώσεις. Στην περίπτωση των προσομοιώσεων χρειάζεται μοντέλο για την περιγραφή του περιβάλλοντος, όμως αυτό δεν δημιουργεί πρόβλημα, καθώς δεν χρειάζεται να είναι γνωστή η ακριβής κατανομή για τη συνάρτηση μετάβασης που περιγράφει το περιβάλλον.

Μία βασική εκδοχή είναι ο αλγόριθμος first-visit Monte Carlo. Ο αλγόριθμος αυτός αρχικοποιεί την τιμή της συνάρτησης αξίας-κατάστασης αυθαίρετα και χρησιμοποιεί μία λίστα για κάθε κατάσταση, η οποία είναι αρχικά κενή. Σε κάθε επανάληψη παράγεται μία ακολουθία από καταστάσεις, δράσεις και ανταμοιβές, σύμφωνα με μία προσομοίωση. Στη συνέχεια, για την πρώτη φορά που παρατηρείται κάθε κατάσταση υπολογίζεται η αναμενόμενη ανταμοιβή (expected return) και αποθηκεύεται στη λίστα που αντιστοιχεί στην κατάσταση αυτή και η τιμή της συνάρτησης αξίας-κατάστασης ανανεώνεται σύμφωνα με τον μέσο όρο των ανταμοιβών που είναι αποθηκευμένες στη λίστα. Για τον αποδοτικότερο υπολογισμό της αναμενόμενης ανταμοιβής, ο υπολογισμός ξεκινάει από την τελική στιγμή, δίνοντας αρχική τιμή $G = 0$ και ανανεώνεται σε κάθε βήμα προς την αρχική κατάσταση ως $G \leftarrow \gamma G + R_{t+1}$.

Μία παραλλαγή αποτελεί η μέθοδος every-visit στην οποία κάθε εμφάνιση της κατάστασης στην ακολουθία του επεισοδίου χρησιμοποιείται για την εκτίμηση της αναμενόμενης αμοιβής. Οι μέθοδοι Monte-Carlo έχουν επίσης το πλεονέκτημα ότι μπορούν να χρησιμοποιηθούν με ανάλογο τρόπο και για την εκτίμηση της συνάρτησης αξίας-δράσης (action-value function), η οποία είναι χρησιμότερη για την περίπτωση που δεν είναι γνωστό κάποιο μοντέλο του περιβάλλοντος. Επίσης, στους αλγόριθμους Monte-Carlo τα δείγματα μπορεί να παράγονται από διαφορετική πολιτική από αυτή που μαθαίνεται, δηλαδή ανήκουν στην κατηγορία off-policy. Αυτό είναι ιδιαίτερα χρήσιμο, γιατί μπορεί να εκμεταλλευτεί για την εξερεύνηση του χώρου λύσεων. Ιδιαίτερα, όταν ο στόχος είναι η εκμάθηση της συνάρτησης αξίας-δράσης, λόγω του μεγάλου πλήθους του πεδίου ορισμού υπάρχει ο κίνδυνος κάποιες περιοχές να μείνουν ανεξερεύνητες.

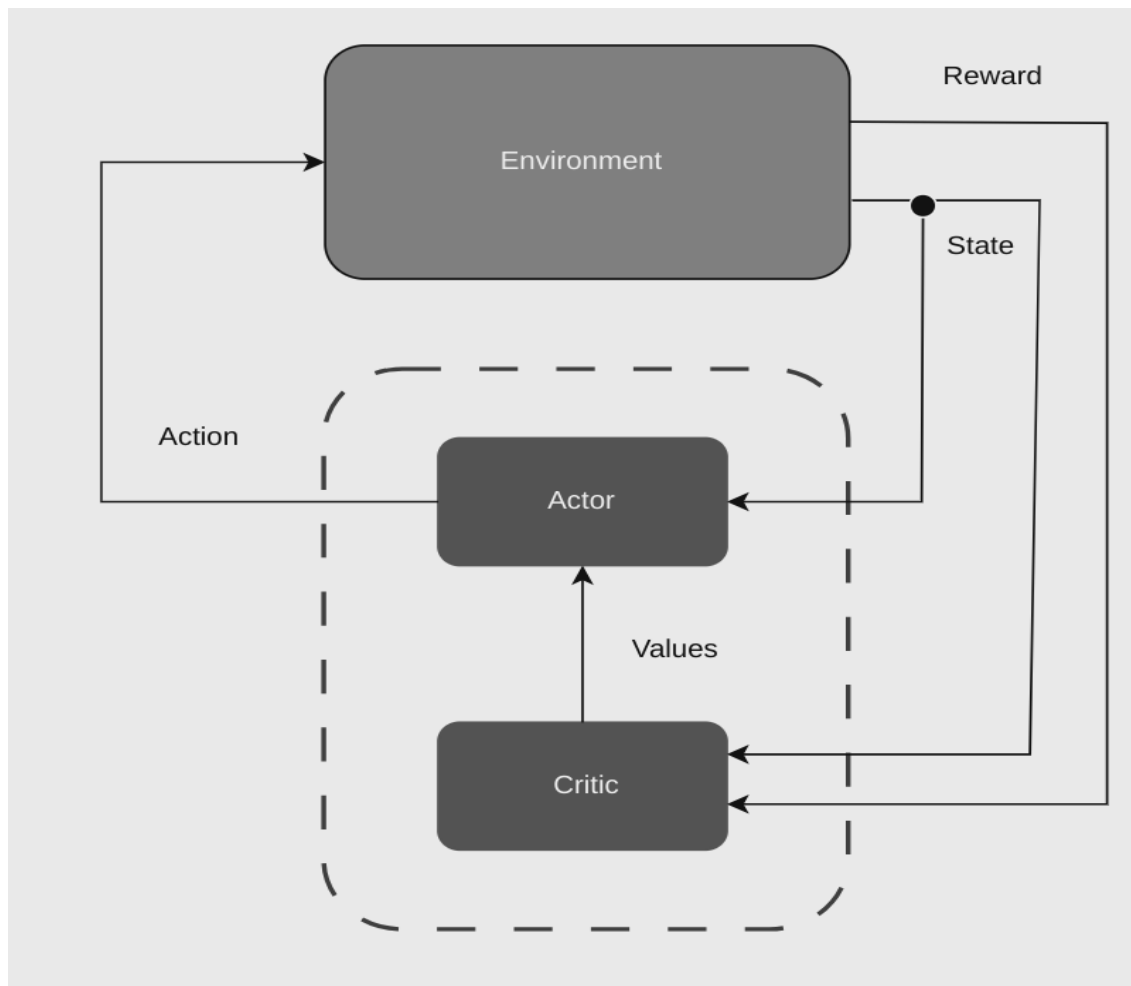
3.5.4 Policy Gradient

Μέθοδος Actor-Critic

Οι μέθοδοι της οικογένειας Actor-Critic βασίζονται σε έναν κριτή (critic), ο οποίος χρησιμοποιείται για να εκτιμήσει την αξία της τρέχουσας πολιτικής, και έναν δράστη (actor), ο οποίος αντιστοιχεί στην πολιτική που έχει μάθει τη δεδομένη στιγμή.

Οι μέθοδοι αυτοί έχουν παρόμοια λογική με αυτή του αλγόριθμου policy iteration υπό την έννοια ότι σε κάθε βήμα χρησιμοποιείται αξιολόγηση της πολιτικής (policy evaluation) και

στη συνέχεια βελτίωσή της (policy improvement). Όμως, η πλήρης αξιολόγηση της πολιτικής δεν είναι δυνατή, όταν η εκπαίδευση γίνεται μέσω δειγμάτων ή χρησιμοποιείται προσέγγιση συνάρτησης. Γι' αυτό και γίνεται ενημέρωση της πολιτικής πριν την πλήρη αξιολόγηση της πολιτικής αυτής[29].



Σχήμα 3.3: Αρχιτεκτονική Actor-Critic

Ο κριτής (critic) χρησιμοποιεί κάποια μέθοδο προσέγγισης συνάρτησης (function approximation) με στόχο να μάθει τη συνάρτηση αξίας-δράσης για την τρέχουσα πολιτική. Η εκτίμηση αυτή βασίζεται σε δείγματα που προκύπτουν από αλληλεπίδραση με το περιβάλλον. Η πολιτική που χρησιμοποιείται για την παραγωγή δειγμάτων (behavior policy) δεν χρειάζεται να είναι ίδια με την πολιτική που μαθαίνεται (target policy). Πολλές φορές χρησιμοποιείται μία ντετερμινιστική πολιτική για την εκμάθηση, ενώ για την δημιουργία δειγμάτων χρησιμοποιείται μία στοχαστική, για να είναι δυνατή η εξερεύνηση. Όμως, η στοχαστική αυτή πολιτική πρέπει να έχει κάποια συσχέτιση με την πολιτική-στόχο, ώστε να ενθαρρύνεται η εξερεύνηση των σημαντικότερων σημείων του χώρου αναζήτησης.

Η πολιτική-στόχος αντιπροσωπεύεται μέσω του δράστη (actor). Ο δράστης χρησιμοποιεί τη συνάρτηση αξίας που έχει μάθει ο κριτής, με σκοπό την εύρεση της βέλτιστης πολιτικής. Ένας τρόπος για να επιλεγεί καλύτερη πολιτική είναι η επιλογή με άπληστο τρόπο, σύμφωνα με την εκτίμηση του critic. Μία εναλλακτική είναι η χρήση ενός βήματος gradient ascent

σε μία συνάρτηση που μετράει την επίδοση της τρέχουσας πολιτικής χρησιμοποιώντας ως χώρο αναζήτησης μία συγκεκριμένη οικογένεια πολιτικών.

Οι μέθοδοι Actor-Critic ανήκουν στην κατηγορία *generalized policy iteration*, στην οποία ανήκουν οι αλγόριθμοι που ενημερώνουν την πολιτική χωρίς να την αξιολογούν πλήρως. Για αυτή την κατηγορία αλγορίθμων, δεν είναι δεδομένη η σύγκλιση στη βέλτιστη τιμή, όπως ισχύει για τον αλγόριθμο *policy iteration*. Έτσι, είναι πιθανό ένα βήμα του αλγορίθμου να οδηγεί σε χειρότερη πολιτική. Γι' αυτό και δεν είναι δεδομένο ότι η τελική πολιτική είναι η βέλτιστη, οπότε πρέπει να αποθηκεύονται ενδιάμεσα μοντέλα και να αξιολογούνται τελικά[29].

Αλγόριθμος Deep Deterministic Policy Gradient (DDPG)

Ο αλγόριθμος DDPG είναι ένας actor-critic, model-free, off-policy αλγόριθμος ενισχυτικής μάθησης. Ο αλγόριθμος αυτό χρησιμοποιεί βαθιά νευρωνικά δίκτυα για την προσέγγιση συναρτήσεων και βασίζεται στο θεώρημα Policy Gradient, με σκοπό την εκμάθηση μίας ντετερμινιστικής πολιτικής. Η εκπαίδευση πραγματοποιείται μέσω δειγμάτων που προκύπτουν μέσω προσομοιώσεων, γι' αυτό και ο περιορισμός στις ντετερμινιστικές πολιτικές μπορεί να βοηθήσει σημαντικά την εκπαίδευση, καθώς η πολιτική είναι συνάρτηση μόνο των καταστάσεων και όχι των δράσεων, οπότε περιορίζεται ο χώρος αναζήτησης και η λύση μπορεί να βρεθεί ταχύτερα.

Θεωρώντας ένα πρόβλημα ενισχυτικής μάθησης, το οποίο περιγράφεται από μία Μαρκοβιανή Διαδικασία Απόφασης με χώρο καταστάσεων $\mathcal{S}$ και χώρο δράσεων $\mathcal{A}$, τότε αν με θ συμβολίζονται οι παράμετροι μέσω των οποίων γίνεται η προσέγγιση της πολιτικής π_θ η κλίση της επίδοσης της πολιτικής (performance gradient) δίνεται από τη σχέση :

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) da ds \\ &= \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)]\end{aligned}\quad (3.14)$$

Όπου $\rho^\pi(s) = \int_{\mathcal{S}} \sum_{t=1}^{+\infty} \gamma^{t-1} p_1(s') p(s' \rightarrow s, t, \pi) ds$ είναι η μειωμένη κατανομή κατάστασης (discounted state distribution) με $p_1(s)$ να δηλώνει την πιθανότητα η s να είναι η αρχική κατάσταση και $p(s' \rightarrow s, t, \pi)$ την πιθανότητα ξεκινώντας από την κατάσταση s' και ακολουθώντας την πολιτική π να εμφανιστεί η κατάσταση s μετά από t χρονικές στιγμές. Αυτό είναι το Stochastic Policy Gradient Theorem και σε αυτό στηρίζονται οι αλγόριθμοι Policy Gradient για τη βελτίωση της πολιτικής που μαθαίνει ο actor [30].

Το σημαντικότερο συμπέρασμα του θεωρήματος αυτού είναι ότι η κλίση της επίδοσης της πολιτικής (policy gradient) δεν εξαρτάται από την κλίση της κατανομής καταστάσεων (state distribution), παρά το γεγονός ότι η κατανομή πιθανότητας εξαρτάται από την πολιτική που ακολουθείται. Έτσι, μπορούν να χρησιμοποιηθούν δείγματα για να εκτιμηθεί η αναμενόμενη αμοιβή.

Οι αλγόριθμοι Policy Gradient αποτελούν την κυριότερη κατηγορία αλγορίθμων για προβλήματα με συνεχείς χώρους δράσεων. Ο λόγος είναι ότι η άπληστη επιλογή βέλτιστης δράσης σε συνεχείς χώρους απαιτεί την επίλυση ενός προβλήματος βελτιστοποίησης σε κάθε βήμα βελτίωσης της πολιτικής. Αντίθετα, η κλίση policy gradient δίνει άμεσα την κατεύθυνση προς την οποία βελτιώνεται η πολιτική.

Ο αλγόριθμος DDPG αποτελεί έναν από τα σημαντικότερα παραδείγματα αυτής της κατηγορίας. Το θεώρημα Policy Gradient για νιτερμινιστικές πολιτικές παίρνει τη μορφή:

$$\begin{aligned}\nabla_{\partial^\mu} J &\approx \mathbb{E}_{s_t \sim \rho^\beta} \left[\nabla_{\partial^\mu} Q(s, a | \partial^\theta) |_{s=s_t, a=\mu(s_t | \partial^\mu)} \right] \\ &= \mathbb{E}_{s_t \sim \rho^\beta} \left[\nabla_a Q(s, a | \partial^\theta) |_{s=s_t, a=\mu(s_t)} \nabla_{\partial^\mu} \mu(s | \partial^\mu) |_{s=s_t} \right]\end{aligned}\quad (3.15)$$

Στον αλγόριθμο DDPG τόσο η προσέγγιση της συνάρτησης του critic όσο και του actor γίνονται μέσω νευρωνικών δικτύων. Όμως, οι αλγόριθμοι βελτιστοποίησης για την εκπαίδευση των νευρωνικών δικτύων που χρησιμοποιούνται θεωρούν ότι τα δείγματα είναι ανεξάρτητα και προέρχονται από την ίδια κατανομή, το οποίο δεν ισχύει για συνεχόμενα δείγματα από ένα επεισόδιο. Γ' αυτόν τον λόγο χρησιμοποιείται μνήμη, η οποία ονομάζεται replay buffer. Πρόκειται για μνήμη σταθερού μεγέθους στην οποία αποθηκεύεται κάθε μετάβαση που προκύπτει από την προσομοίωση με τη μορφή πλειάδας (s_t, a_t, r_t, s_{t+1}) . Όταν γεμίσει, το παλαιότερο δείγμα αντικαθίσταται από το νέο. Για την εκπαίδευση χρησιμοποιείται ένα μικρό σύνολο (mini batch) από δείγματα που επιλέγονται τυχαία από τον replay buffer. Η χρήση mini batch για την εκπαίδευση χρησιμεύει και στην αξιοποίηση βελτιστοποιήσεων από το υλικό για την αποδοτικότερη εκμάθηση [31].

Επιπρόσθετα, χρησιμοποιούνται αντίγραφα των critic και actor, τα οποία ονομάζονται target critic και target actor αντίστοιχα. Οι παράμετροι αυτών των δικτύων μεταβάλλονται με σημαντικά μικρότερο ρυθμό. Με αυτόν τον τρόπο επιτυγχάνεται η σταθερότητα κατά την εκπαίδευση, παρά το γεγονός ότι επιβραδύνεται η ταχύτητα μάθησης. Σχετικά με την εξερεύνηση, εφόσον ο αλγόριθμος DDPG είναι off-policy, τα δείγματα δημιουργούνται από διαφορετική πολιτική από την $\mu(s_t | \partial^\mu)$. Επομένως, για την παραγωγή δειγμάτων προστίθεται θόρυβος στην πολιτική αυτή, ώστε να επιτυγχάνεται η εξερεύνηση.

Ο αλγόριθμος αρχικά θέτει τις παραμέτρους των target networks Q', μ' στις ίδιες τιμές με τα δίκτυα του critic Q και του actor μ . Στη συνέχεια, επαναλαμβάνονται επεισόδια προσομοιώσεων του αλγορίθμου, όπου σε κάθε χρονική στιγμή επιλέγεται μία δράση $a_t = \mu(s_t | \partial^\mu) + \mathcal{N}_t$, όπου $\mathcal{N}_t$ τυχαίος θόρυβος. Από την εκτέλεση αυτής της δράσης προκύπτει μία πλειάδα (s_t, a_t, r_t, a_{t+1}) , η οποία αποθηκεύεται στον replay buffer. Έπειτα, δειγματοληπτείται τυχαία ένα σύνολο από N μεταβάσεις από τον replay buffer. Η ενημέρωση του critic γίνεται ελαχιστοποιώντας τη συνάρτηση απώλειας (loss):

$$L = \frac{1}{N} \sum_i (r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1} | \partial^{\mu'}) | \partial^{Q'}) - Q(s_i, a_i | \partial^Q))^2 \quad (3.16)$$

Η βελτίωση της πολιτικής γίνεται με χρήση του θεωρήματος Policy Gradient, το οποίο υπολογίζεται κατά προσέγγιση βάσει των δειγμάτων από τη σχέση:

$$\nabla_{\partial^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a | \partial^Q) |_{s=s_i, a=\mu(s_i)} \nabla_{\partial^\mu} \mu(s | \partial^\mu) |_{s_i} \quad (3.17)$$

Τέλος, τα δίκτυα target ανανεώνονται σύμφωνα με τις παρακάτω σχέσεις:

$$\theta^Q \leftarrow \tau \theta^Q + (1 - \tau) \bar{\theta}^Q \quad (3.18)$$

$$\theta^\mu \leftarrow \tau \theta^\mu + (1 - \tau) \bar{\theta}^\mu \quad (3.19)$$

Σημειώνεται ότι η τιμή του τ είναι θετική αλλά πολύ μικρότερη από 1.

3.6 Ενισχυτική Μάθηση Πολλαπλών Πρακτόρων

Η ενισχυτική μάθηση πολλαπλών πρακτόρων (Multi-Agent Reinforcement Learning - MARL) αποτελεί γενίκευση της ενισχυτικής μάθησης για την εκπαίδευση δύο ή περισσότερων πρακτόρων. Στα πλαίσια του MARL κάθε πράκτορας αποτελεί διακριτή οντότητα, η οποία αλληλεπιδρά με το κοινό για όλους περιβάλλον. Η βάση της περιγραφής των προβλημάτων MARL είναι το Μαρκοβιανό παίγνιο (Markov game).

Μία βασική διάκριση των περιπτώσεων εφαρμογής γίνεται βάσει των στόχων των πρακτόρων. Αν όλοι οι πράκτορες έχουν έναν κοινό στόχο, τότε γίνεται λόγος για συνεργατικό (cooperative) περιβάλλον. Από την άλλη, εμφανίζονται περιπτώσεις, όπου οι πράκτορες έχουν αντικρουόμενους στόχους, οπότε γίνεται λόγος για ανταγωνιστικό (competitive) περιβάλλον. Τέλος, υπάρχουν εφαρμογές που οι πράκτορες έχουν έναν κοινό στόχο, ενώ ο κάθε πράκτορας έχει ακόμη έναν ατομικό στόχο. Σε αυτή την περίπτωση, το περιβάλλον θεωρείται ανάμεικτο (mixed).

Σε περιβάλλοντα πολλαπλών πρακτόρων εγείρονται ζητήματα σχετικά με τις δυνατότητες επικοινωνίας και συντονισμού μεταξύ των πρακτόρων. Αυτά τα ζητήματα αποτελούν σημαντικούς παράγοντες για τον σχεδιασμό της λύσης του προβλήματος. Ανάλογα με τις πληροφορίες που μοιράζονται οι πράκτορες κατά την εκπαίδευση και κατά την εκτέλεση ορίζονται τρεις κατηγορίες για την ενισχυτική μάθηση πολλαπλών πρακτόρων.

3.6.1 Κατηγορίες MARL

Distributed Training Decentralized Execution (DTDE)

Πρόκειται για την περίπτωση που οι πράκτορες δεν βασίζονται στην ανταλλαγή πληροφορίας μεταξύ τους, αλλά μαθαίνουν ανεξάρτητα ο ένας από τον άλλο (independent learners). Χαρακτηριστικά παραδείγματα αυτής της κατηγορίας είναι η εφαρμογή αλγορίθμων ενισχυτικής μάθησης ενός πράκτορα για τον κάθε έναν πράκτορα ανεξάρτητα.

Επειδή ο κάθε πράκτορας δεν έχει γνώση για την κατάσταση ή τις δράσεις των υπολοίπων, το περιβάλλον δεν είναι στάσιμο από την οπτική του. Αυτό αποτελεί το κυριότερο μειονέκτημα αυτής της κατηγορίας, καθώς οι μεταβολές, λόγω των υπόλοιπων πρακτόρων, δεν γίνονται αντιληπτές, όμως επηρεάζουν την πολιτική που μαθαίνεται. Επίσης, απαιτούν τη δημιουργία δειγμάτων για εκπαίδευση για τον κάθε πράκτορα ξεχωριστά με αποτέλεσμα να δυσχεραίνεται η κλιμακωσιμότητα τους. Τέλος, η εφαρμογή αυτών των μεθόδων σε στοχαστικά περιβάλλοντα παρουσιάζει ιδιαίτερες δυσκολίες, καθώς δεν είναι δυνατή η διάκριση μεταξύ της στοχαστικότητας του περιβάλλοντος και της στοχαστικότητας που χρησιμοποιείται για εξερεύνηση της βέλτιστης πολιτικής[32].

Centralized Training Centralized Execution (CTCE)

Στην κατηγορία αυτή θεωρείται ότι υπάρχει μία κεντρική οντότητα, η οποία μαθαίνει μία απο κοινού πολιτική για όλους τους πράκτορες. Για την εκμάθηση αυτής της πολιτικής χρησιμοποιούνται άμεσα οι αλγόριθμοι ενισχυτικής μάθησης ενός πράκτορα. Η εφαρμογή τέτοιων μεθόδων είναι περιορισμένη, καθώς για να εφαρμοστούν πρέπει να είναι δυνατή η άμεση και απρόσκοπτη επικοινωνία μεταξύ των πρακτόρων. Ακόμη, ο χώρος δράσεων-καταστάσεων είναι εκθετικός ως προς το πλήθος των πρακτόρων, με αποτέλεσμα τέτοιοι αλγόριθμοι να μην είναι κλιμακώσιμοι.

Centralized Training Decentralized Execution (CTDE)

Σε αυτή την κατηγορία επιτρέπεται η ανταλλαγή πληροφορίας μεταξύ των πρακτόρων μόνο κατά τη διαδικασία της εκπαίδευσης. Αυτή η κοινή πληροφορία αξιοποιείται προκειμένου να διευκολυνθεί η διαδικασία και να επιταχυνθεί η εκμάθηση. Επίσης, περιορίζονται προβλήματα μη στασιμότητας που παρατηρούνται στην κατανεμημένη εκπαίδευση. Τα παραπάνω καθιστούν αυτή την κατηγορία ως την κατηγορία αιχμής για το πεδίο της ενισχυτικής μάθησης πολλαπλών πρακτόρων.

Οι μέθοδοι αυτές παρουσιάζουν περαιτέρω πλεονεκτήματα. Σε περιπτώσεις που ο στόχος των πρακτόρων είναι κοινός, μπορεί να γίνει διαμοιρασμός των παραμέτρων προς εκμάθηση προκειμένου να επιταχυνθεί η διαδικασία. Επίσης, με χρήση τεχνικών αποσύνθεσης του συνολικού προβλήματος σε μικρότερα υποπροβλήματα μπορεί να βελτιωθεί η κλιμακωσιμότητα τους. Τέλος, λόγω της διαφοροποίησης μεταξύ εκπαίδευσης και εκτέλεσης, μπορεί κατά την εκπαίδευση να χρησιμοποιηθεί περαιτέρω πληροφορία, η οποία δεν θα είναι διαθέσιμη κατά την εκπαίδευση προκειμένου να επιτευχθούν καλύτερα αποτελέσματα[32].

3.6.2 Multi-Agent Deep Deterministic Policy Gradient (MADDPG)

Στην παρούσα εργασία θα εφαρμοστεί ο αλγόριθμος Multi-Agent Deep Deterministic Policy Gradient (MADDPG). Πρόκειται για έναν αλγόριθμο ενισχυτικής μάθησης πολλαπλών πρακτόρων, ο οποίος είναι model-free, online και off-policy. Ο αλγόριθμος MADDPG χρησιμοποιεί το σχήμα της κεντριοποιημένης εκπαίδευσης με αποκεντρωμένη εκτέλεση (centralized training with decentralized execution). Για κάθε πράκτορα ακολουθείται η αρχιτεκτονική actor-critic, ώστε να είναι δυνατή η χρήση πλεονάζουσας πληροφορίας μόνο κατά την εκπαίδευση[6].

Θεωρώντας ένα Markov game, το οποίο περιγράφει το ζητούμενο πρόβλημα, ο κάθε πράκτορας αποκτά πληροφορία μόνο για ένα υποσύνολο της συνολικής κατάστασης, το οποίο ονομάζεται παρατήρηση (observation) και συμβολίζεται με O_i . Το σύνολο αυτό είναι διαφορετικό για κάθε πράκτορα. Ο σκοπός του αλγορίθμου είναι η εύρεση της βέλτιστης ντετερμινιστικής πολιτικής $\mu_i : O_i \rightarrow \mathcal{A}_i$, η οποία πρέπει να έχει ως πεδίο ορισμού την παρατήρηση του πράκτορα, ώστε να μπορεί να χρησιμοποιηθεί άμεσα κατά την εκτέλεση.

Αναφορικά με τον critic που χρησιμοποιείται για την εκτίμηση της αναμενόμενης ανταμοιβής της τρέχουσας πολιτικής ορίζεται ως συνάρτηση της συνολικής κατάστασης S και των δράσεων κάθε πράκτορα. Αυτό δε δημιουργεί προβλήματα στην αποκεντρωμένη εκτέλεση,

καθώς ο critic χρησιμοποιείται μόνο κατά την εκπαίδευση. Έτσι, η συνολική κατάσταση περιέχει το σύνολο των παρατηρήσεων όλων των πρακτόρων και πιθανώς επιπλέον πληροφορία για το περιβάλλον. Με αυτόν τον τρόπο το περιβάλλον φαίνεται στάσιμο από την οπτική του κάθε πράκτορα, καθώς είναι γνωστή η δράση και των υπόλοιπων πρακτόρων. Τέλος, σημειώνεται ότι ο κάθε πράκτορας χρησιμοποιεί το δικό του δίκτυο για τον critic.

Σχετικά με το θεώρημα Policy Gradient για την περίπτωση των πολλαπλών πρακτόρων, αυτό έχει την ακόλουθη μορφή. Αν $J(\theta_i) = \mathbb{E}[\mathcal{R}_i]$ είναι η αναμενόμενη ανταμοιβή, με θ_i τις παραμέτρους που ορίζουν την πολιτική του πράκτορα i , τότε η κλίση της υπολογίζεται ως:

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s \sim p^\mu, a_i \sim \pi_i} [\nabla_{\theta_i} \log \pi_i(a_i | o_i) Q_i^\pi(s, a_1, \dots, a_N)] \quad (3.20)$$

Ο αλγόριθμος MADDPG βασίζεται στην αλληλεπίδραση των πρακτόρων με το περιβάλλον για να παραχθούν δείγματα. Για να γίνεται επαρκής εξερεύνηση προστίθεται θόρυβος στη δράση που προκύπτει από την τρέχουσα πολιτική. Όπως συμβαίνει και στον αλγόριθμο DDPG, τα δείγματα που παράγονται δεν χρησιμοποιούνται απευθείας για την εκπαίδευση, αλλά αποθηκεύονται σε έναν replay buffer.

Αντίστοιχα, με τον αλγόριθμο DDPG χρησιμοποιούνται target critic networks και target actor networks, τα οποία είναι αντίγραφα των main networks, αλλά ανανεώνονται με μικρότερο ρυθμό. Ο λόγος που χρησιμοποιούνται είναι να δώσουν σταθερότητα στην εκπαίδευση, λόγω του μικρού ρυθμού μεταβολής τους.

Σε κάθε εκτέλεση του αλγορίθμου εκπαίδευσης, η εκπαίδευση γίνεται με minibatch από δείγματα που επιλέγονται ομοιόμορφα από τον replay buffer. Αφού ο κάθε πράκτορας επιλέξει το minibatch από S δείγματα, τα οποία έχουν τη μορφή (s, a, r, s') , υπολογίζει για κάθε δείγμα j την τιμή:

$$y^j = r_i^j + \gamma \cdot Q_i^{\mu'}(s'^j, a_1^j, \dots, a_N^j) \Big|_{a_k' = \mu_k'(o_k^j)}$$

Όπου Q_i' το target critic network και μ_i' το target actor network. Βάσει αυτών των τιμών οι παράμετροι του critic ανανεώνονται, ώστε να ελαχιστοποιείται η συνάρτηση απώλειας:

$$\mathcal{L}(\theta_i) = \frac{1}{S} \sum_j (y^j - Q_i^\mu(s^j, a_1^j, \dots, a_N^j))^2$$

Αντίστοιχα, ανανεώνονται οι παράμετροι του actor εκτελώντας ένα βήμα του αλγορίθμου gradient ascent, υπολογίζοντας το gradient, βάσει του θεωρήματος Policy Gradient, ως:

$$\nabla_{\theta_i} J \approx \frac{1}{S} \sum_j [\nabla_{\theta_i} \mu_i(o_i^j) \cdot \nabla_{a_i} Q_i^\mu(s^j, a_1^j, \dots, a_i, \dots, a_N^j)] \Big|_{a_i = \mu_i(o_i^j)}$$

Τέλος, οι παράμετροι των δικτύων target για κάθε πράκτορα, ανανεώνονται σύμφωνα με τη σχέση:

$$\theta_i' \leftarrow \tau \cdot \theta_i + (1 - \tau) \cdot \theta_i'$$

Όπου τ θετική σταθερά, πολύ μικρότερη από τη μονάδα.

Ο αλγόριθμος που χρησιμοποιείται για την εκπαίδευση των πρακτόρων σε μορφή ψευδοκώδικα :

ΑΛΓΟΡΙΘΜΟΣ 3.1: MADDPG training

for each agent i **do**

Sample a random minibatch of S samples (s, a, r, s') from the replay buffer

for each sample j in the minibatch **do**

Compute $y^j = r_i^j + \gamma \cdot \mathcal{G}_i^{\mu_i}(s'^j, a_1', \dots, a_N') \Big|_{a_k' = \mu_k'(o_k^j)}$

end for

Update the critic by executing one step of the gradient descent algorithm to minimize the loss:

$$\mathcal{L}(\partial_i) = \frac{1}{S} \sum_j \left(y^j - \mathcal{G}_i^{\mu}(s^j, a_1^j, \dots, a_N^j) \right)^2$$

Update the actor by executing one step of the gradient ascent algorithm to maximize the expected return, when computing the gradient of the expected return as:

$$\nabla_{\partial_i} J = \frac{1}{S} \sum_j [\nabla_{\partial_i} \mu_i(o_i^j) \cdot \nabla_{a_i} \mathcal{G}_i^{\mu}(s^j, a_1^j, \dots, a_i, \dots, a_N^j)] \Big|_{a_i = \mu_i(o_i^j)}$$

Update the weights of the target networks:

$$\partial_i' \leftarrow \tau \cdot \partial_i + (1 - \tau) \cdot \partial_i'$$

end for

Κεφάλαιο 4

Μοντελοποίηση προβλήματος

Στο κεφάλαιο αυτό το πρόβλημα εξισορρόπησης του παραγόμενου φορτίου με το ζητούμενο, δηλαδή η ισχύς που παρατηρείται στο PCC να είναι ίση με την ισχύ που καθορίζει το πλάνο διανομής (dispatch plan) τυποποιείται ως πρόβλημα μαθηματικής βελτιστοποίησης. Συγκεκριμένα, παρουσιάζεται η αρχική μορφή του προβλήματος, καθώς και εναλλακτικές που στοχεύουν στη δίκαιη χρήση των μπαταριών χωρίς να υπονομεύεται σημαντικά ο αρχικός στόχος. Μεταξύ των τρόπων που χρησιμοποιούνται με σκοπό την εισαγωγή της δικαιοσύνης στο πρόβλημα είναι και η μέθοδος των τροχιών και γι' αυτό τον λόγο παρουσιάζονται μέθοδοι, ώστε να εισαχθούν τα κριτήρια δικαιοσύνης κατά τον υπολογισμό των τροχιών.

Στη συνέχεια του κεφαλαίου το πρόβλημα του ελέγχου μοντελοποιείται ως πρόβλημα ενισχυτικής μάθησης πολλαπλών πρακτόρων. Συγκεκριμένα, παρουσιάζονται δύο εναλλακτικές περιγραφές για την κατάσταση κάθε μπαταρίας, οι οποίες κρίνονται κατάλληλες για το πρόβλημα. Έπειτα, παρουσιάζονται συναρτήσεις ανταμοιβής, οι οποίες στηρίζονται στις εναλλακτικές περιγραφές που παρουσιάζονται στην ενότητα 4.1 και στοχεύουν στη δίκαιη χρήση μεταξύ των μπαταριών.

4.1 Μαθηματική Περιγραφή Προβλήματος

Θεωρώντας ένα δίκτυο ενέργειας, το οποίο αποτελείται κόμβους όπου ο κάθε ένας θεωρείται ότι συνδέεται με μία μπαταρία, συνολικού πλήθους N_B , το ζητούμενο είναι η ισχύς στο σημείο σύνδεσης με το δίκτυο μεταφοράς, δηλαδή στο PCC να ισούται με την ισχύ που καθορίζει το πλάνο διανομής (dispatch plan). Ο έλεγχος λαμβάνει χώρα για ένα χρονικό παράθυρο, όπου αποφάσεις λαμβάνονται τις χρονικές στιγμές $\{k, k + 1, \dots, T - 1\}$, ενώ ως T θεωρείται η τελική χρονική στιγμή του παραθύρου ελέγχου. Μεταξύ των παραπάνω χρονικών στιγμών μεσολαβεί χρόνος Δt . Ο παραπάνω στόχος μοντελοποιείται μαθηματικά μέσω της ελαχιστοποίησης του τετραγωνικού σφάλματος μεταξύ των δύο τιμών. Ακόμη, η κάθε μπαταρία έχει ένα σύνολο από περιορισμούς τόσο σχετικά με την κατάσταση της ενέργειάς της (SoE_i) όσο και για την ισχύ που παράγει ή απορροφά (B_i). Τέλος, μπορεί να υπάρχουν περιορισμοί σχετικά με το άθροισμα των SoE συγκεκριμένων μπαταριών σε καθορισμένες χρονικές στιγμές, οι οποίοι προκύπτουν από ευρύτερους στόχους του δικτύου.

Στην εργασία, κατά σύμβαση με $B_i > 0$ συμβολίζεται η φόρτιση της μπαταρίας, ενώ με $P_{disp}(t), P_{PCC}(t) > 0$ η ισχύς εισέρχεται προς το δίκτυο μέσω του συστήματος μεταφοράς.

Τέλος, με $L(t) > 0$ δηλώνεται ότι τα μη ελεγχόμενα στοιχεία του δικτύου καταναλώνουν ενέργεια.

Όλα τα παραπάνω, οδηγούν στο σχεδιασμό του παρακάτω προβλήματος κυρτής βελτιστοποίησης:

$$\min_{P_{PCC}(t), B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} (P_{PCC}(t) - P_{disp}(t))^2 \quad (4.1)$$

$$\text{subject to } P_{PCC}(t) = \sum_{i=1}^{N_B} B_i(t) + L(t) \quad (4.2)$$

$$SoE_i(t+1) = SoE_i(t) + B_i(t)\Delta t \quad (4.3)$$

$$B_i^{min} \leq B_i(t) \leq B_i^{max} \quad (4.4)$$

$$SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \quad (4.5)$$

$$SoE_j^{min, t_j} \leq \sum_{i=1}^{N_B} a_{ji} SoE_i(t_j) \leq SoE_j^{max, t_j} \quad (4.6)$$

$$\text{for } i \in \{1, 2, \dots, N_B\}, t \in \{k, \dots, T-1\}, j \in \{1, \dots, N_C\}$$

Η συνάρτηση βελτιστοποίησης (4.1) αποτελεί το συνολικό τετραγωνικό σφάλμα της απόκλισης των τιμών της ισχύος P_{PCC} στο PCC και της ισχύος P_{disp} που ορίζει το πλάνο διανομής. Η ελαχιστοποίηση της παραπάνω συνάρτησης αντιστοιχεί στον στόχο του να ακολουθείται το πλάνο διανομής. Η εξίσωση (4.2) αποτελεί τη σχέση μέσω της οποίας υπολογίζεται η παρατηρούμενη ισχύς στο PCC. Η σχέση αυτή προκύπτει για την περίπτωση, όπου δεν υπάρχουν απώλειες ενέργειας, όπως υποτίθεται στην παρούσα εργασία. Η εξίσωση (4.3) δηλώνει τη σχέση μέσω της οποίας μεταβάλλεται η κατάσταση της ενέργειας κάθε μπαταρίας λόγω της δράσης της. Η σχέση (4.4) και η σχέση (4.5) δηλώνουν τους περιορισμούς αναφορικά με την ελάχιστη και μέγιστη δράση κάθε μπαταρίας και αναφορικά με την ελάχιστη και τη μέγιστη τιμή για την κατάσταση της ενέργειας κάθε μπαταρίας αντίστοιχα. Σημειώνεται ότι ισχύει $B_i^{min} \leq 0 \leq B_i^{max}$ για κάθε μπαταρία.

Για το πρόβλημα δίνεται η αρχική συνθήκη για το SoE κάθε μπαταρίας $SoE_i(k)$. Η τιμή αυτή είναι σημαντική, καθώς η βέλτιστη λύση εξαρτάται από αυτή λόγω του περιορισμού (4.3), για $t = k$. Επίσης, η τιμή αυτή πρέπει να ικανοποιεί τον περιορισμό $SoE_i^{min} \leq SoE_i(k) \leq SoE_i^{max}$ για κάθε μπαταρία, παρ' όλο που δεν αναφέρεται στους περιορισμούς του προβλήματος, καθώς δεν μπορεί να ελεγχθεί στο χρονικό παράθυρο που εξετάζει το συγκεκριμένο πρόβλημα.

Τέλος, η εξίσωση (4.6) συμβολίζει τους συνδυαστικούς περιορισμούς που αφορούν το άθροισμα ενός καθορισμένου συνόλου μπαταριών να βρίσκεται σε συγκεκριμένο εύρος μια ορισμένη χρονική στιγμή. Σχετικά με το a_{ji} , μπορεί να έχει την τιμή 1, οπότε η μπαταρία i συμμετέχει στον περιορισμό j , ή την τιμή 0 διαφορετικά. Πιο συγκεκριμένα, υπάρχουν N_C τέτοιοι περιορισμοί, όπου ο j -οστός περιορισμός ορίζει ότι τη χρονική στιγμή $t_j \in \{k+1, \dots, T\}$ το άθροισμα για των SoE των μπαταριών που ανήκουν στο σύνολο $\{i \mid a_{ji} \neq 1\}$. Ακόμη, για τους συνδυαστικούς περιορισμούς πρέπει να ισχύει για κάθε $j \in \{1, \dots, N_C\}$ η σχέση $\sum_{i=1}^{N_B} a_{ji} SoE_i^{min} \leq SoE_j^{min, t_j} \leq SoE_j^{max, t_j} \leq \sum_{i=1}^{N_B} a_{ji} SoE_i^{max}$, ώστε να υπάρχουν εφικτές λύσεις στο πρόβλημα, αλλά και να είναι χρήσιμος ο περιορισμός. Για κάποιο περιορισμό j μπορεί να

ισχύει η σχέση $\sum_{i=1}^{N_B} a_{ji} = 1$. Σε αυτή την περίπτωση ο περιορισμός αφορά μόνο μία μπαταρία. Στην παρούσα εργασία ο όρος συνδυαστικός περιορισμός θα περιλαμβάνει "καταχρηστικά" και τέτοιου είδους περιορισμούς.

Το προηγούμενο πρόβλημα δεν λαμβάνει υπόψιν ζητήματα δίκαιης χρήσης μεταξύ των μπαταριών. Ένας τρόπος, ώστε να εισαχθεί ένα κριτήριο δικαιοσύνης σε ένα πρόβλημα βελτιστοποίησης είναι μέσω της κατάλληλης αλλαγής της αντικειμενικής συνάρτησης, ώστε δίκαιες λύσεις με ίδιο κόστος να έχουν μικρότερη αντικειμενική τιμή. Εναλλακτική επιλογή αποτελεί η προσθήκη ενός κριτηρίου δικαιοσύνης ως ρητό περιορισμό (hard constraint) στο πρόβλημα βελτιστοποίησης, εφόσον είναι εφικτό και δεν επηρεάζει την κυρτότητα του συνολικού προβλήματος.

Όπως έχει παρουσιαστεί προηγουμένως οι τροχιές αποτελούν μία μέθοδο, η οποία μπορεί να χρησιμοποιηθεί για τον σκοπό αυτό. Ο λόγος είναι ότι υπολογίζονται offline από ένα σύνολο σεναρίων, οπότε είναι ευκολότερο να εισαχθεί ένα κριτήριο. Στην επόμενη ενότητα, προτείνονται μέθοδοι που μπορούν να χρησιμοποιηθούν για να παραχθούν δίκαιες τροχιές αξιοποιώντας τη μέθοδο που παρουσιάζεται στο [10].

Εφόσον έχουν υπολογιστεί οι τροχιές, τότε η αντικειμενική συνάρτηση θα πρέπει να ενθαρρύνει τις μπαταρίες να ακολουθούν αυτές τις τροχιές. Μία τέτοια αντικειμενική συνάρτηση είναι η

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} \left[\left(P_{PCC}(t) - P_{disp}(t) \right)^2 + \frac{a_p}{w(t)} \cdot \sum_{i=1}^{N_B} C_{i,t}(B_i(t)|SoE_i(t)) \right] \quad (4.7)$$

Με συνάρτηση κόστους μπαταρίας $C_{i,t}(B_i(t)|SoE_i(t))$ την ακόλουθη

$$C_{i,t}(B_i|SoE_i) = \frac{(B_i(t) - B_i^*(SoE_i(t), t))^2}{4B_i^{max}} \quad (4.8)$$

Ενώ ισχύουν οι περιορισμοί (4.2) - (4.6). Με $B_i^*(SoE_i(t), t)$ συμβολίζεται η προτίμηση της μπαταρίας, η οποία ορίζεται στην [ενότητα 2.4](#) και υπολογίζεται σύμφωνα με την εξίσωση (2.14). Με a_p συμβολίζεται ο συντελεστής του κόστους μπαταρίας, ο οποίος έχει μία σταθερή τιμή.

Η μέθοδος αυτή ακολουθεί τη λογική της μεθόδου που παρουσιάζεται στην [ενότητα 2.4](#). Όμως υπάρχει η σημαντική διαφορά που προκύπτει λόγω της χρονικής εξάρτησης. Συγκεκριμένα, η τιμή του $B_i^*(SoE_i(t), t)$ επηρεάζεται από τις αποφάσεις για τις προηγούμενες χρονικές στιγμές, επομένως τα θεωρητικά αποτελέσματα που παρουσιάστηκαν στην ενότητα παύουν να είναι ακριβή για την περίπτωση αυτή.

Η μορφή του κόστους μπαταρίας μπορεί να αναχθεί σε αυτή της εξίσωσης (2.11), αγνοώντας την άεργο ισχύ. Συγκεκριμένα, η σταθερά c αντικαθίσταται από τον όρο $\left(\frac{B_i^*(SoE_i(t), t)}{2\sqrt{B_i^{max}}} \right)^2$, ο οποίος για τη μέθοδο που οι αποφάσεις λαμβάνονται ανεξάρτητα αποτελεί σταθερά. Επίσης, αντί του S_i χρησιμοποιείται η τιμή του B_i^{max} , καθώς ικανοποιεί τις ίδιες υποθέσεις με το S_i , ενώ υπάρχει στις υπάρχουσες υποθέσεις του προβλήματος της εργασίας, οπότε χρειάζεται να θεωρηθούν περισσότερα μεγέθη για το δίκτυο.

Η επιλογή τα βάρη να μην πολλαπλασιάζουν το κόστος δικτύου, αλλά ο αντίστροφος τους να πολλαπλασιάζει το κόστος μπαταρίας δικαιολογείται, ώστε να εμφανίζεται στη συνάρτηση

ο όρος που αφορά το συνολικό κόστος δικτύου αυτούσιος. Έτσι, η απόκλιση από το πλάνο διανομής ανεξάρτητα από τη χρονική στιγμή που γίνεται. Αντίθετα, η απόκλιση της μπαταρίας από την προτίμηση μπορεί να έχει διαφορετική βαρύτητα σε διαφορετικές χρονικές στιγμές. Συγκεκριμένα, αν κάποια χρονική στιγμή η τιμή του $w(t)$ παίρνει μεγάλη τιμή, τότε το κόστος μπαταρίας πολλαπλασιάζεται με τον συντελεστή $\frac{ap}{w(t)}$, ο οποίος θα έχει σχετικά μικρή τιμή, άρα θα δωθεί μικρότερη βαρύτητα στον στόχο της μπαταρίας. Αντίστοιχα, για μικρές τιμές του $w(t)$ δίνεται μεγαλύτερη βαρύτητα στο να ακολουθούν οι μπαταρίες τις τροχιές τους.

Οι παραπάνω μέθοδοι εισάγουν έμμεσα τη δίκαιη χρήση μέσω των τροχιών και δεν είναι προφανές ότι οι λύσεις που προκύπτουν είναι απαραίτητα δίκαιες. Έτσι, θα ήταν χρήσιμο το κόστος της άδικης χρήσης να ενσωματωθεί απευθείας στη συνάρτηση βελτιστοποίησης. Στην ενότητα 2.5.1 ορίστηκε ότι η δίκαιη χρήση μεταξύ δύο μπαταριών i, j είναι να ικανοποιούν τη σχέση $\frac{B_i(t)}{B_i^{max}} = \frac{B_j(t)}{B_j^{max}}$. Έτσι, το κόστος για την άδικη χρήση μεταξύ δύο μπαταριών για συγκεκριμένη χρονική στιγμή μπορεί να οριστεί ως η παράσταση $\left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}}\right)^2$. Το κόστος αυτό δηλώνει την τετραγωνική απόκλιση από την απολύτως δίκαιη χρήση, σύμφωνα με το κριτήριο δικαιοσύνης της υποενότητας 2.5.1. Το συνολικό κόστος μπορεί να χρησιμοποιηθεί απευθείας στη συνάρτηση βελτιστοποίησης, η οποία περιλαμβάνει επίσης το συνολικό κόστος δικτύου και χρησιμοποιώντας την κατάλληλη βαρύτητα στον κάθε στόχο, ώστε να εξισορροπείται η έμφαση που δίνεται στον καθένα σχηματίζεται το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} \left[(P_{PCC}(t) - P_{disp}(t))^2 + w_f \cdot \sum_{i=1}^{N_B} \sum_{j=1}^{N_B} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2 \right] \quad (4.9)$$

Ενώ ισχύουν οι περιορισμοί (4.2) - (4.6).

Η παραπάνω συνάρτηση προσθέτει, μέσω του όρου $\left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}}\right)^2$, ως κόστος το τετραγωνικό σφάλμα της απόκλισης από το κριτήριο δικαιοσύνης. Η τιμή του βάρους w_f καθορίζει την έμφαση που θα δοθεί στο κριτήριο δικαιοσύνης σε σχέση με το κόστος δικτύου. Μικρά βάρη ενθαρρύνουν λύσεις με μικρό κόστος, όχι όμως δίκαιες, ενώ μεγάλα βάρη οδηγούν σε λύσεις δίκαιες, οι οποίες μπορεί όμως να έχουν σημαντικά αυξημένο κόστος δικτύου.

Αν ο σκοπός είναι όσο το δυνατόν πιο δίκαιη χρήση, ακόμη κι αν έχει σημαντικά μεγαλύτερο κόστος, τότε μπορεί να χρησιμοποιηθεί το πρόβλημα (4.1) και με περιορισμούς (4.2) - (4.6), καθώς και τον περιορισμό

$$\frac{B_i(t)}{B_i^{max}} = \frac{B_j(t)}{B_j^{max}} \quad (4.10)$$

$$\forall i, j \in \{1, \dots, N_B\}, t \in \{k, \dots, T\}$$

Η χρήση του παραπάνω περιορισμού περιορίζει σημαντικά τις εφικτές λύσεις και μπορεί να οδηγήσει σε αρκετές περιπτώσεις σε ανέφικτα προβλήματα, οπότε χρειάζεται προσοχή όταν επιλέγεται να χρησιμοποιηθεί.

4.2 Σχεδιασμός Δίκαιων Τροχιών

Μία πρώτη μέθοδος υπολογισμού τροχιών αποτελεί η μέθοδος της ενότητας 2.3. Συγκεκριμένα, ως συνάρτηση κόστους θα χρησιμοποιούταν η συνάρτηση (2.1) και θα ίσχυαν οι περιορισμοί (2.2), (2.6) και (4.3) - (4.6). Στο παράρτημα Α αποδεικνύεται ότι για τη γενική περίπτωση η παραπάνω μέθοδος δεν εγγυάται τη δίκαιη χρήση μεταξύ των μπαταριών.

4.2.1 Δίκαιες τροχιές βάσει των περιορισμών ισχύος

Στην παρούσα ενότητα παρουσιάζονται μέθοδοι, οι οποίες παράγουν τροχιές, οι οποίες είναι δίκαιες βάσει του κριτηρίου της ενότητας 2.5.1. Ο πρώτος τρόπος είναι να εισαχθεί στη συνάρτηση κόστους το τετραγωνικό σφάλμα για την απόκλιση των ισχύων των μπαταριών από τη δίκαιη χρήση. Δηλαδή ως συνάρτηση βελτιστοποίησης να χρησιμοποιηθεί η εξής:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} \left\{ \frac{1}{D} \sum_{d=1}^D [P_{PCC,d}(t) - P_{disp}(t)]^2 + w_f \cdot \sum_{i=1}^{N_B} \sum_{j=1}^{N_B} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2 \right\} \quad (4.11)$$

Υπό τους περιορισμούς (2.2), (2.6) και (4.3) - (4.6).

Η τιμή του βάρους w_f είναι αυτή που καθορίζει αν θα δοθεί περισσότερη έμφαση στον στόχο του να ακολουθείται το πλάνο διανομής (dispatch plan) ή τον στόχο της δίκαιης χρήσης. Μεγάλες τιμές του w_f δίνουν μεγαλύτερη βαρύτητα στη δικαιοσύνη, όμως η λύση μπορεί να έχει σημαντικά αυξημένο κόστος αναφορικά με το κόστος του δικτύου (grid objective). Αντίθετα, για μικρές τιμές του w_f υπάρχει ο κίνδυνος οι τροχιές που προκύπτουν να μην είναι δίκαιες. Σε αυτή την περίπτωση όμως, αναμένεται η λύση να είναι κοντά στη βέλτιστη για τον στόχο του συνολικού δικτύου.

Ένας διαφορετικός τρόπος είναι να χρησιμοποιηθεί η συνθήκη δικαιοσύνης στους περιορισμούς του προβλήματος. Δηλαδή, ως συνάρτηση κόστους ορίζεται η:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} \left\{ \frac{1}{D} \sum_{d=1}^D [P_{PCC,d}(t) - P_{disp}(t)]^2 \right\} \quad (4.12)$$

Στο σύνολο που ορίζεται από τους περιορισμούς (2.2), (2.6), (4.3) - (4.6) και

$$\frac{B_i(t)}{B_i^{max}} = \frac{B_j(t)}{B_j^{max}} \quad (4.13)$$

$$\forall t \in \{k, \dots, T-1\}, i, j \in \{1, \dots, N_B\}$$

Λόγω της χρήσης αυστηρού περιορισμού (hard constraint) υπάρχει ο κίνδυνος το παραπάνω πρόβλημα βελτιστοποίησης να μην έχει εφικτές λύσεις, οπότε η συγκεκριμένη μέθοδος δεν μπορεί να χρησιμοποιηθεί σε κάθε περίπτωση.

Μέσω του περιορισμού (4.13) διασφαλίζεται η μέγιστη δυνατή δικαιοσύνης. Έτσι, η μέθοδος αυτή υπολογίζει τη βέλτιστη τροχιά ως προς το κόστος δικτύου από το σύνολο των δικαιότερων δυνατών τροχιών.

Τέλος, όπως αποδεικνύεται στο παράρτημα Α μπορεί να υπάρχουν πολλές βέλτιστες τροχιές. Σε αυτή την περίπτωση, για να επιλεγεί η δικαιότερη από το σύνολο των βέλτιστων

λύσεων μπορεί να χρησιμοποιεί η παρακάτω μέθοδος.

Έστω $B_i^{opt}(t)$ η τιμή της μεταβλητής $B_i(t)$, όπως προκύπτει από υπολογισμό τροχιών σύμφωνα με την αντικειμενική συνάρτηση (2.1), δηλαδή μία λύση η οποία δεν είναι δίκαιη. Για την εύρεση της δικαιότερης λύσης επιλύεται το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} \sum_{i=1}^{N_B} \sum_{j=1}^{N_B} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2 \quad (4.14)$$

$$\text{subject to } \sum_{i=1}^{N_B} B_i(t) = \sum_{i=1}^{N_B} B_i^{opt}(t) \quad (4.15)$$

$$B_i^{min} \leq B_i(t) \leq B_i^{max} \quad (4.16)$$

$$SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \quad (4.17)$$

$$SoE_j^{min, t_j} \leq \sum_{i=1}^{N_B} a_{ji} SoE_i(t_j) \leq SoE_j^{max, t_j} \quad (4.18)$$

$$\text{for } i \in \{1, 2, \dots, N_B\}, t \in \{k, \dots, T-1\}, j \in \{1, \dots, N_C\}$$

Η λύση αυτή θεωρείται η δικαιότερη, καθώς ελαχιστοποιεί τη συνολική τετραγωνική απόκλιση από την απολύτως δίκαιη χρήση, σύμφωνα με το αντίστοιχο κριτήριο. Η λύση αυτή εξακολουθεί να είναι βέλτιστη αναφορικά με το συνολικό κόστος του δικτύου, καθώς ο περιορισμός που περιγράφεται μέσω της εξίσωσης (4.15) εξασφαλίζει ότι η τιμή του $P_{PCC}(t)$ θα είναι αυτή που ήταν και στη βέλτιστη λύση, άρα το συνολικό άθροισμα $\sum_{t=k}^{T-1} (P_{PCC}(t) - P_{disp}(t))^2$ παραμένει το ελάχιστο.

Σημειώνεται αν κάθε μπαταρία συμμετέχει σε περιορισμούς της μορφής (4.18) πάντα με το ίδιο σύνολο μπαταριών, δηλαδή αν για κάθε i , αν $a_{ji} = a_{li} = 1$, τότε για κάθε μπαταρία m ισχύει $a_{jm} = a_{lm}$, τότε μπορεί να χρησιμοποιηθεί η παρακάτω περιγραφή του προβλήματος:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} \sum_{i=1}^{N_B} \sum_{l \in \mathcal{J}: a_{li}=a_{ji}=1} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_l(t)}{B_l^{max}} \right)^2 \quad (4.19)$$

$$\text{subject to } \sum_{i=1}^{N_B} a_{ji} B_i(t) = \sum_{i=1}^{N_B} a_{ji} B_i^{opt}(t) \quad (4.20)$$

$$B_i^{min} \leq B_i(t) \leq B_i^{max} \quad (4.21)$$

$$SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \quad (4.22)$$

$$\text{for } i \in \{1, 2, \dots, N_B\}, t \in \{k, \dots, T-1\}, j \in \{1, \dots, N_C\}$$

Θέτοντας $B_i(t) = B_i^{opt}(t)$ για κάθε μπαταρία που ισχύει ότι $\sum_{j=1}^{N_C} a_{ji} = 0$, δηλαδή για κάθε μπαταρία που δεν συμμετέχει σε κανέναν συνδυαστικό περιορισμό.

Στην περίπτωση αυτή η σχέση (4.20) εξασφαλίζει εκτός από την ελαχιστοποίηση του κόστους δικτύου και την ικανοποίηση των συνδυαστικών περιορισμών, όπως προκύπτει από την περιγραφή στο [παράρτημα Α](#), ενώ η εφικτότητα του δεύτερου προβλήματος διασφαλίζεται από το γεγονός ότι λύση του πρώτου αποτελεί εφικτή λύση του δεύτερου.

4.2.2 Δίκαιες τροχιές βάσει της τρέχουσας Κατάστασης της Ενέργειας

Αντίστοιχα, για κριτήριο δικαιοσύνης μπορεί να χρησιμοποιηθεί αυτό που περιγράφεται στην ενότητα 2.5.2. Καθώς απαιτείται να είναι γνωστό αν πρέπει να γίνει φόρτιση ή εκφόρτιση σε κάθε χρονικό βήμα, ο υπολογισμός γίνεται σε δύο βήματα.

Το αρχικό βήμα είναι η επίλυση του προβλήματος (2.1) με τους περιορισμούς (2.2), (2.6) και (4.3) - (4.6). Μέσω του βήματος αυτού μαθαίνεται μία βέλτιστη λύση $B_i^{opt}(t)$, η οποία αξιοποιείται για να καθοριστεί αν πρέπει να φορτίσουν ή εκφορτίσουν οι μπαταρίες τη χρονική στιγμή t . Συγκεκριμένα, αν ισχύει $\sum_{i=1}^{N_B} B_i^{opt}(t) > 0$, τότε οι μπαταρίες πρέπει να φορτίσουν τη χρονική στιγμή t , ενώ αν είναι $\sum_{i=1}^{N_B} B_i^{opt}(t) < 0$, τότε πρέπει να εκφορτίσουν. Αν το άθροισμα είναι 0, τότε οι μπαταρίες παραμένουν ανενεργές.

Σχετικά με το δεύτερο βήμα, αυτό αφορά τον υπολογισμό της δικαιότερης λύσης. Μία πρώτη σκέψη θα ήταν να χρησιμοποιηθεί η έννοια του proportional fairness, για να μοιραστεί η συνολική ισχύς $P_{ref} = \sum_{i=1}^{N_B} B_i^{opt}(t)$, που πρέπει να προσφερθεί ή απορροφηθεί προκειμένου να ελαχιστοποιηθεί το κόστος δικτύου. Λόγω των περιορισμών είναι πιθανό η απόλυτη δικαιοσύνη να μην είναι εφικτή, γι' αυτό και χρησιμοποιείται ο αλγόριθμος 4.1, ο οποίος ακολουθεί τη λογική του max-min fairness χρησιμοποιώντας βάρη, για να προσεγγίσει το κριτήριο δικαιοσύνης 2.5.2.

Πιο συγκεκριμένα, αν τη χρονική στιγμή t πρέπει να γίνει εκφόρτιση αρχικά τίθεται $P_{ref} = \sum_{i=1}^{N_B} B_i^{opt}(t)$, δηλαδή η συνολική ισχύ που θα παρήγαγαν οι μπαταρίες σύμφωνα με τη λύση του προβλήματος βελτιστοποίησης. Για κάθε μπαταρία, αρχικά τίθεται η τιμή

$$B_i(t) = \max \left(\frac{SoE_i(t) - SoE_i^{min}}{\sum_{j=1}^{N_B} (SoE_j(t) - SoE_j^{min})} \cdot P_{ref}, B_i^{min} \right) \quad (4.23)$$

Η ισχύς αναφοράς ανανεώνεται με την τιμή $P'_{ref} = P_{ref} - \sum_{i=1}^{N_B} B_i^{opt}(t)$ και επαναλαμβάνεται η ίδια διαδικασία, αγνοώντας από το άθροισμα τις μπαταρίες που έχουν την οριακή τιμή, προσθέτοντας σε κάθε βήμα στην προηγούμενη τιμή μέχρι το P_{ref} να μηδενιστεί. Αντίστοιχα για τις χρονικές στιγμές που πρέπει να φορτίσουν οι μπαταρίες για την ενημέρωση της τιμής της ισχύος χρησιμοποιείται η σχέση

$$B_i(t) = \min \left(\frac{SoE_i^{max} - SoE_i(t)}{\sum_{j=1}^{N_B} (SoE_j^{max} - SoE_j(t))} \cdot P_{ref}, B_i^{max} \right) \quad (4.24)$$

ενώ τα υπόλοιπα βήματα είναι ίδια.

Η παραπάνω μέθοδος διασφαλίζει ότι ικανοποιούνται οι συνδυαστικοί περιορισμοί μόνο αν σε αυτούς συμμετέχουν όλες οι μπαταρίες. Η απόδειξη παρουσιάζεται στο [Παράρτημα Α](#). Για τις υπόλοιπες περιπτώσεις πρέπει να ελέγχεται αν επιτυγχάνουν τον στόχο να ικανοποιείται ο περιορισμός. Επίσης, δεν εξασφαλίζεται ότι ικανοποιείται ο περιορισμός $SoE_i^{min} \leq SoE_i(t) \leq SoE_i^{max}$, για κάθε μπαταρία, όμως ισχύει ότι $SoE_i^{min} + B_i^{min} \cdot \Delta t \leq SoE_i(t) \leq SoE_i^{max} + B_i^{max} \cdot \Delta t$. Επομένως, μια λύση είναι να σχεδιαστούν υποβέλτιστες τροχιές για εκείνες τις περιπτώσεις που δεν ικανοποιείται ο περιορισμός, οι οποίες θα χρησιμοποιούν για τα όρια του SoE τις τιμές $SoE_i^{min'} = SoE_i^{min} - B_i^{min} \cdot \Delta t$ και $SoE_i^{max'} = SoE_i^{max} - B_i^{max} \cdot \Delta t$. Εναλλακτικά, μπορεί να χρησιμοποιηθεί η επόμενη μέθοδος η οποία δουλεύει σε κάθε πε-

ΑΛΓΟΡΙΘΜΟΣ 4.1: Υπολογισμός δίκαιων τροχιών βάσει του κριτηρίου 2.5.2

```

for ( $t \leftarrow k \cdot t < T \cdot + + t$ ) do
     $B_i(t) \leftarrow 0, \forall i = \{1, \dots, N_B\}$ 
     $d_i(t) \leftarrow 1, \forall i = \{1, \dots, N_B\}$ 
     $P_{ref} \leftarrow \sum_{i=1}^{N_B} B_i^{OPT}(t)$ 
    while  $P_{ref} < 0$  do
        for all batteries with  $d_i \neq 0$  do
             $B_i(t) \leftarrow \max \left( B_i(t) + \frac{SoE_i(t) - SoE_i^{min}}{\sum_{j=1}^{N_B} d_i \cdot (SoE_j(t) - SoE_j^{min})} \cdot P_{ref}, B_i^{min} \right)$ 
            if  $B_i(t) = B_i^{min}$  then
                 $d_i \leftarrow 0$ 
            end if
        end for
         $P_{ref} \leftarrow \sum_{i=1}^{N_B} B_i^{OPT}(t) - \sum_{i=1}^{N_B} B_i(t)$ 
    end while
    while  $P_{ref} > 0$  do
        for all batteries with  $d_i \neq 0$  do
             $B_i(t) \leftarrow \min \left( B_i(t) + \frac{SoE_i^{max} - SoE_i(t)}{\sum_{j=1}^{N_B} (SoE_j^{max} - SoE_j(t))} \cdot P_{ref}, B_i^{max} \right)$ 
            if  $B_i(t) = B_i^{max}$  then
                 $d_i \leftarrow 0$ 
            end if
        end for
         $P_{ref} \leftarrow \sum_{i=1}^{N_B} B_i^{OPT}(t) - \sum_{i=1}^{N_B} B_i(t)$ 
    end while
end for

```

ρίπτωση.

Μια εναλλακτική είναι η ελαχιστοποίηση ενός μέσου τετραγωνικού σφάλματος βάσει των σχέσεων που περιγράφουν το κριτήριο δικαιοσύνης. Για τη μέθοδο αυτή οι αποφάσεις θα λαμβάνονται ανεξάρτητα κάθε χρονική στιγμή, καθώς οι συναρτήσεις που υποθέτουν χρονική εξάρτηση είναι γενικά μη κυρτές για το συγκεκριμένο κριτήριο δικαιοσύνης. Επειδή οι σχέσεις $\frac{B_i(t)}{SoE_i^{max}-SoE_i(t)} = \frac{B_j(t)}{SoE_j^{max}-SoE_j(t)}$, $\frac{B_i(t)}{SoE_i(t)-SoE_i^{min}} = \frac{B_j(t)}{SoE_j(t)-SoE_j^{min}}$ μπορεί να μην ορίζονται για κάποιες χρονικές στιγμές, αφού είναι πιθανό ο παρονομαστής να πάρει την τιμή μηδέν, για το μέσο τετραγωνικό σφάλμα χρησιμοποιείται η παρακάτω εναλλακτική περιγραφή των σχέσεων:

$$B_i(t) \cdot (SoE_j(t) - SoE_j^{min}) - B_j(t) \cdot (SoE_i(t) - SoE_i^{min}) = 0 \quad (4.25)$$

$$B_i(t) \cdot (SoE_j^{max} - SoE_j(t)) - B_j(t) \cdot (SoE_i^{max} - SoE_i(t)) = 0 \quad (4.26)$$

Σχετικά με τις χρονικές στιγμές αυτές διαμερίζονται στα σύνολα $T_1 = \{t | P_{ref}(t) < 0\}$ και $T_2 = \{t | P_{ref}(t) > 0\}$. Επομένως, σειριακά αποφασίζεται η τιμή της ενεργού ισχύος κάθε μπαταρίας επιλύοντας το ακόλουθο πρόβλημα βελτιστοποίησης, αν $t \in T_1$:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{i=1}^{N_B} \sum_{n=1}^{N_B} (B_i(t) \cdot (SoE_n(t) - SoE_n^{min}) - B_n(t) \cdot (SoE_i(t) - SoE_i^{min}))^2 \quad (4.27)$$

$$\text{subject to } \sum_{i=1}^{N_B} B_i(t) = \sum_{i=1}^{N_B} B_i^{opt}(t) \quad (4.28)$$

$$B_i^{min} \leq B_i(t) \leq B_i^{max} \quad (4.29)$$

$$SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \quad (4.30)$$

$$SoE_j^{min, t_j} \leq \sum_{i=1}^{N_B} a_{ji} (SoE_i(t) + B_i^{max} \cdot \Delta t \cdot \Delta N_j(t)) \quad (4.31)$$

$$\sum_{i=1}^{N_B} a_{ji} (SoE_i(t) + B_i^{min} \cdot \Delta t \cdot \Delta N_j(t)) \leq SoE_j^{max, t_j} \quad (4.32)$$

$$\text{for } i \in \{1, 2, \dots, N_B\}, t \in \{k, \dots, T-1\}, j \in \{1, \dots, N_C\}$$

$$\text{Όπου η συνάρτηση } \Delta N_j(t) = \begin{cases} t_j - t & , t \leq t_j \\ T & , \text{else} \end{cases}$$

Δηλαδή δίνει το πλήθος το χρονικών στιγμών έως και τη στιγμή του περιορισμού, αν η χρονική στιγμή προηγείται του περιορισμού, ενώ διαφορετικά δίνει T , δηλαδή το πλήθος όλων των χρονικών στιγμών, ώστε να ικανοποιείται ο περιορισμός.

Ενώ αν είναι $t \in T_2$, επιλύεται το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{i=1}^{N_B} \sum_{n=1}^{N_B} (B_i(t) \cdot (SoE_n^{max} - SoE_n(t)) - B_n(t) \cdot (SoE_i^{max} - SoE_i(t)))^2 \quad (4.33)$$

Υπό τους περιορισμούς (4.28) - (4.32). Η μέθοδος αυτή εγγυάται την ικανοποίηση των περιορισμών σε κάθε περίπτωση. Σε περίπτωση που για κάποιο πρόβλημα προκύπτει ότι δεν υπάρχουν εφικτές λύσεις, τότε ο περιορισμός (4.28) χαλαρώνεται και τίθεται ως ποινή

στην αντικειμενική συνάρτηση προς βελτιστοποίηση, δηλαδή για $t \in T_1$ χρησιμοποιείται η ακόλουθη συνάρτηση:

$$\begin{aligned} \min_{B_1(t), \dots, B_{N_B}(t)} & \sum_{i=1}^{N_B} \sum_{n=1}^{N_B} (B_i(t) \cdot (SoE_n(t) - SoE_n^{min}) - B_n(t) \cdot (SoE_i(t) - SoE_i^{min}))^2 \\ & + w_r \cdot \left(\sum_{i=1}^{N_B} (B_i(t) - B_i^{opt}(t)) \right)^2 \end{aligned}$$

Όπου w_r το βάρος που δίνεται για τον όρο του περιορισμού που χαλαρώθηκε και πρέπει να έχει μεγάλη τιμή.

4.3 Αναπαράσταση της Κατάστασης

Για την εφαρμογή του αλγορίθμου MADDPG χρειάζεται να περιγραφεί το πρόβλημα ως Μαρκοβιανό Παίγνιο (Markov Game). Στην υλοποίηση χρησιμοποιείται ένας πράκτορας για κάθε μπαταρία, άρα το πλήθος των πρακτόρων είναι N_B .

Σχετικά με την αναπαράσταση της κατάστασης κάθε πράκτορα, η Κατάσταση της Ενέργειας (SoE) του οφείλει να περιλαμβάνεται σε αυτή, καθώς είναι απαραίτητη για την ικανοποίηση των περιορισμών. Ακόμη, η χρονική στιγμή t συμπεριλαμβάνεται, καθώς οι συνδυαστικοί περιορισμοί αφορούν μελλοντικές χρονικές στιγμές, οπότε η απόσταση από αυτές πρέπει να είναι γνωστή.

Τέλος, κατά τη διάρκεια του ελέγχου η τιμή του πραγματικού φορτίου που παράγουν οι ΑΠΕ δεν είναι γνωστή τη στιγμή που λαμβάνεται η απόφαση κάθε πράκτορα. Για τον λόγο αυτό πρέπει να χρησιμοποιηθεί μία εκτίμηση αυτής της τιμής. Στην εργασία αυτή για εκτίμηση του μελλοντικού φορτίου $\tilde{L}(t)$ χρησιμοποιείται η τιμή του φορτίου την αμέσως προηγούμενη χρονική στιγμή $L(t-1)$.

Στο επίπεδο ελέγχου πραγματικού χρόνου στο οποίο στοχεύει η παρούσα εργασία πρέπει να αξιοποιείται η πιο πρόσφατη πληροφορία, ώστε να προκύπτουν τα βέλτιστα αποτελέσματα. Επομένως, πληροφορία για την τιμή $\tilde{L}(t)$ οφείλει να περιέχεται στην κατάσταση. Στη συνέχεια, παρουσιάζονται δύο εναλλακτικές αναπαραστάσεις για την κατάσταση κάθε πράκτορα. Η βέλτιστη επιλογή για την αναπαράσταση δεν μπορεί να αποδειχθεί θεωρητικά, οπότε οι δύο αναπαραστάσεις συγκρίνονται στη συνέχεια μέσω υπολογιστικών προσομοιώσεων.

Η πρώτη επιλογή είναι η κατάσταση κάθε πράκτορα να περιγράφεται από την πλειάδα $(SoE_i(t), \tilde{L}(t), t)$. Σε αυτή την περίπτωση, η συνολική κατάσταση του περιβάλλοντος περιγράφεται από την πλειάδα $(SoE_1(t), \dots, SoE_{N_B}(t), \tilde{L}(t), t)$. Η συνάρτηση μετάβασης μεταξύ των καταστάσεων περιγράφεται από τις σχέσεις:

$$SoE_i(t) \leftarrow SoE_i(t) + B_i(t) \cdot \Delta t \quad (4.34)$$

$$\tilde{L}(t) \leftarrow L(t) \quad (4.35)$$

$$t \leftarrow t + 1 \quad (4.36)$$

Σχετικά με τη σχέση (4.35) δεν είναι συνάρτηση της προηγούμενης κατάστασης και των δράσεων, αλλά περιγράφεται από την κατανομή πιθανότητας να παρατηρηθεί η τιμή $L(t)$ τη

χρονική στιγμή t .

Η εναλλακτική αναπαράσταση είναι μέσω της πλειάδας $(SoE_i(t), P_{PCC}(t-1), B_i(t-1), t)$. Δηλαδή αντί του φορτίου από τα μη ελεγχόμενα στοιχεία χρησιμοποιείται η συνολική παρατηρούμενη ισχύς στο PCC κατά την προηγούμενη χρονική στιγμή, ενώ έχει προστεθεί στην κατάσταση και η ενεργός ισχύς της ίδιας μπαταρία για την προηγούμενη χρονική στιγμή. Μέσω της περιγραφής αυτής στη συνολική κατάσταση εξακολουθεί να υπάρχει στη συνολική κατάσταση η τιμή του $\tilde{L}(t)$. Αυτό ισχύει, καθώς η συνολική κατάσταση περιγράφεται από την πλειάδα $(SoE_1(t), B_1(t-1), \dots, SoE_{N_B}(t), B_{N_B}(t-1), P_{PCC}(t-1), t)$. Όμως, ισχύει η σχέση $\tilde{L}(t) = P_{PCC}(t-1) - \sum_{i=1}^{N_B} B_i(t-1)$. Σχετικά με τη συνάρτηση μετάβασης, αυτή περιγράφεται από τις σχέσεις:

$$SoE_i(t) \leftarrow SoE_i(t) + B_i(t) \cdot \Delta t \quad (4.37)$$

$$P_{PCC}(t-1) \leftarrow L(t) + \sum_{j=1}^{N_B} B_j(t) \quad (4.38)$$

$$B_i(t-1) \leftarrow B_i(t) \quad (4.39)$$

$$t \leftarrow t + 1 \quad (4.40)$$

Η σχέση (4.38) ορίζεται μέσω της κατανομής πιθανότητας για την τιμή του $L(t)$ τη δεδομένη χρονική στιγμή, επομένως ούτε σε αυτή την αναπαράσταση η συνάρτηση μετάβασης είναι ντετερμινιστική συνάρτηση.

Σημειώνεται ότι τόσο η τιμή $\tilde{L}(t)$, όσο και η τιμή $P_{PCC}(t-1)$ δεν είναι τοπικά γνωστές στον πράκτορα, οπότε προϋποτίθεται η ανάγκη επικοινωνίας με το PCC προκειμένου να ενημερωθεί για την τιμή. Έτσι, καμία δεν εμφανίζει κάποιο πρακτικό πλεονέκτημα έναντι της άλλης. Επομένως, η σύγκριση αναμένεται να γίνει μέσω προσομοιώσεων. Το ερώτημα που εγείρει η σύγκριση των μεθόδων είναι αν η αφηρημένη πληροφορία σχετικά με την ισχύ των υπόλοιπων μπαταριών μαζί με το φορτίο των μη ελεγχόμενων στοιχείων, όπως ορίζεται στη δεύτερη περίπτωση, είναι χρήσιμη για τον έλεγχο.

4.4 Συνάρτηση Ανταμοιβής βάσει Τροχιών

Η συνάρτηση ανταμοιβής αποτελεί κρίσιμη επιλογή για την επιτυχία της εφαρμογής των αλγορίθμων ενισχυτικής μάθησης. Μία συνάρτηση ανταμοιβής ιδανικά συμπεριλαμβάνει τους άμεσους στόχους του προβλήματος δίνοντας την κατάλληλη βαρύτητα στον καθένα, ενώ δίνει και επαρκή πληροφορία για μελλοντικούς στόχους.

Μία δυνατή επιλογή για τη συνάρτηση αμοιβής, η οποία αντιστοιχεί στο πρόβλημα βελτιστοποίησης (4.7) και ακολουθεί τη λογική της μεθόδου της ενότητας 2.4 είναι η ακόλουθη:

$$r_i(t) = \left(\tilde{L}(t) + \sum_{j=1}^{N_B} B_j(t) - P_{disp}(t) \right)^2 + \frac{a_p}{w_m(t)} \frac{(B_i(t) - B_i^*(SoE_i(t), t))^2}{4B_i^{max}} + \psi_i(t) + \Lambda_i(t) \quad (4.41)$$

Όπου $\tilde{L}(t) = L(t-1)$ είναι η εκτίμηση για το μη ελεγχόμενο φορτίο, a_p ο συντελεστής του κόστους μπαταρίας (battery objective), ενώ $w_m(t)$ η βαρύτητα του κόστους δικτύου

(grid objective). Με $\psi(t)$ συμβολίζεται η συνάρτηση ποινής (penalty function), η οποία χρησιμοποιείται ως χαλάρωση του περιορισμού (soft constraint) για τα όρια της κατάστασης της ενέργειας της μπαταρίας. Αντίστοιχα, η $\Lambda_i(t)$ αντιστοιχεί στη συνάρτηση ποινής για τους συνδυαστικούς περιορισμούς σε καθορισμένες χρονικές στιγμές.

Η εισαγωγή των τροχιών στο πρόβλημα έχει ως κύριο στόχο να επιτυγχάνεται η δικαιοσύνη μεταξύ της χρήσης των μπαταριών. Ακόμη, είναι χρήσιμη γιατί οι τροχιές ικανοποιούν τους περιορισμούς του προβλήματος, οπότε λειτουργούν ως proxy για μελλοντικούς περιορισμούς. Επιπλέον, πιθανό όφελος της χρήσης των τροχιών είναι ότι δίνουν την κατεύθυνση για να εξερευνηθεί μία περιοχή, στην οποία βρίσκονται καλές λύσεις. Ο λόγος που η περιοχή γύρω των τροχιών αναμένεται να είναι καλή είναι ότι τα σενάρια που χρησιμοποιηθεί για τον υπολογισμό των τροχιών και τα σενάρια που χρησιμοποιούνται για την εκπαίδευση των πρακτόρων ενισχυτικής μάθησης επιλέγονται από την ίδια κατανομή.

Η επιλογή των βαρών βασίζεται στην πρόταση της εργασίας [2]. Σημειώνεται ότι υπάρχουν σημαντικές διαφορές μεταξύ των δύο προβλημάτων. Στο παρόν πρόβλημα υπάρχει χρονική εξάρτηση, λόγω του $B_i^*(SoE_i(t), t)$ και του περιορισμού $SoE_i(t+1) = SoE_i(t) + B_i(t) \cdot \Delta t$, ενώ αντίθετα η μέθοδος της ενότητας 2.4 επιλέγει δράσεις ανεξάρτητα σε κάθε χρονική στιγμή, χωρίς να λαμβάνει υπόψη τη μελλοντική επίδραση αυτής. Ακόμη, η εργασία [2] δεν έχει περαιτέρω περιορισμούς πέραν της μέγιστης και ελάχιστης ισχύος. Αντίθετα στην εργασία αυτή θεωρούνται περαιτέρω περιορισμοί, οι οποίοι εφόσον εισέρχονται στη συνάρτηση βελτιστοποίησης ως συναρτήσεις ποινής αναμένεται να επηρεάζουν το σημείο της βέλτιστης λύσης.

Παρά τις διαφορές που υπάρχουν θα χρησιμοποιηθεί η μέθοδος εκείνη ως ευρετική για τον υπολογισμό των βαρών και στη συνέχεια θα εξεταστεί μέσω προσομοιώσεων αν λειτουργεί και σε αυτή την περίπτωση. Ακολουθεί μία περιγραφή της μεθόδου προσαρμοσμένη στις υποθέσεις της παρούσας εργασίας και δίνεται διαισθητική ερμηνεία των εννοιών, η οποία εξηγεί τους λόγους που κρίνεται χρήσιμη η εφαρμογή της μεθόδου, παρά τις διαφοροποιήσεις.

Η μέθοδος αυτή λαμβάνει κάθε χρονική στιγμή ανεξάρτητα αποφάσεις επιλύοντας το πρόβλημα βελτιστοποίησης:

$$\begin{aligned} \min_{B_1(t), \dots, B_{N_B}(t)} J(P_{PCC}(t)) + \sum_{i=1}^{N_B} \frac{a_P}{w_m(t)} \frac{(B_i(t) - B_i^*(SoE_i(t), t))}{4B_i^{max}} \\ \text{subject to } B_i(t) \in [B_i^{min}, B_i^{max}], \forall i \in \{1, \dots, N_B\} \end{aligned} \quad (4.42)$$

Όπου:

- $J(P_{PCC}(t)) = (P_{PCC}(t) - P_{disp}(t))^2$
- $P_{PCC}(t) = \tilde{L}(t) + \sum_{i=1}^{N_B} B_i(t)$

Για την επιλογή του βάρους $w_m(t)$ η επιλογή γίνεται βάσει ενός στόχου για τη λύση του προβλήματος βελτιστοποίησης (4.42). Αν με $P_{opt}(t)$ συμβολίζεται η ισχύς στο PCC για τη βέλτιστη λύση, τότε ο στόχος είναι η απόκλιση αυτού από το dispatch plan να βρισκείται

εντός ενός ορίου d_{max} , δηλαδή:

$$|P_{opt}(t) - P_{disp}(t)| \leq d_{max} \quad (4.43)$$

Για τον καθορισμό της τιμής των βαρών χρησιμοποιείται το παρακάτω λήμμα, το οποίο αποδεικνύεται για τη γενική περίπτωση στο [2]:

Λήμμα 4.1. Υποθέτοντας ότι η συνάρτηση για το grid objective $J(P_{PCC}(t))$ είναι αυστηρά κυρτή ως προς P_{PCC} , ο συντελεστής του battery cost είναι φραγμένος, ώστε $\alpha_P \leq 1$ και για την ενεργό ισχύ της μπαταρίας οι περιορισμοί έχουν τη μορφή $B_i(t) \in [B_i^{min}, B_i^{max}]$, για κάθε $i \in \{1, \dots, N_B\}$. Ορίζονται, ακόμη, οι συναρτήσεις:

$$K_t(w_m, d) = \frac{1}{w_m J'(P_{disp}(t) + d)} \quad (4.44)$$

$$h(K) = \begin{cases} |P_{disp}(t) - P_{inf}(t)| & , K \leq 1 \\ |P_{disp}(t) - P_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}| & , K > 1 \end{cases} \quad (4.45)$$

$$g(w_m, d) = \max\{d, h(K_t(w_m, d))\} \quad (4.46)$$

$$f(w_m(t)) = \min_{d>0} g(w_m(t), d) \quad (4.47)$$

Όπου $J'(P_{PCC}(t))$ η παράγωγος της συνάρτησης $J(P_{PCC}(t))$ ως προς $P_{PCC}(t)$ και $P_{inf}(t)$ η προβολή του $P_{disp}(t)$ στο διάστημα $[\tilde{L}(t) + \sum_{i=1}^{N_B} B_i^{min}, \tilde{L}(t) + \sum_{i=1}^{N_B} B_i^{max}]$. Τότε η συνάρτηση $f(w_m(t))$ είναι άνω φράγμα, για την απόκλιση $|P_{opt}(t) - P_{disp}(t)|$, όπου P_{opt} η βέλτιστη λύση του (4.42) για βάρος w_m .

Στην πραγματικότητα παρατηρούνται οι τιμές ισχύος $P_{real}(t) = P_{opt}(t) + \Delta P$, όπου $\Delta P = L(t) - \tilde{L}(t)$ ορίζεται ως η αβεβαιότητα της μη ελεγχόμενης παραγωγής. Επομένως, ισχύει ότι:

$$\begin{aligned} |P_{real}(t) - P_{disp}(t)| &= |P_{real}(t) - P_{opt}(t) + P_{opt}(t) - P_{disp}(t)| \\ &\leq |P_{real}(t) - P_{opt}(t)| + |P_{opt}(t) - P_{disp}(t)| \\ &\leq |\Delta P|_{max} + f(w_m(t)) \end{aligned}$$

Από λήμμα 4.1 και θέτοντας ως $|\Delta P|_{max}$ τη μέγιστη αβέβαιη παραγωγή ή κατανάλωση, δηλαδή η μέγιστη τιμή του ΔP .

Ορίζεται η συνάρτηση $f_u(w_m(t)) = f(w_m(t)) + |\Delta P|_{max}$, η οποία έχει τις ακόλουθες ιδιότητες:

1. Είναι φθίνουσα ως προς $w_m(t)$
2. Έχει άνω φράγμα $f_{UB} = |P_{disp}(t) - P_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}| + |\Delta P|_{max}$
3. Έχει κάτω φράγμα $f_{LB} = |P_{disp}(t) - P_{inf}(t)| + |\Delta P|_{max}$
4. Αν $P_{disp}(t) \neq P_{inf}(t)$, τότε $f_u(w_m(t)) = f_{LB}$, για κάθε $w_m(t) \geq \frac{1}{|J'(P_{inf}(t))|}$

Στο Παράρτημα Β παρουσιάζεται αναλυτικότερα η συνάρτηση $f(w_m(t))$ και γίνεται άμεσα αντιληπτό, γιατί ισχύουν οι ιδιότητες 1-3. Σχετικά με την τέταρτη ιδιότητα, ακολουθεί η απόδειξη.

Απόδειξη. Περίπτωση 1: Αν $P_{disp}(t) > P_{inf}(t)$. Είναι:

$$\begin{aligned} g(w_m(t), P_{disp}(t) - P_{inf}(t)) &= \max\{P_{disp}(t) - P_{inf}(t), h(K_t(w_m(t), P_{disp}(t) - P_{inf}(t)))\} \\ &= P_{disp}(t) - P_{inf}(t) \end{aligned}$$

Γιατί εφόσον η $J'(P_{PCC}(t))$ είναι γνησίως αύξουσα ισχύει ότι $J'(P_{inf}) > 0$, ενώ επειδή είναι $w_m(t) \geq \frac{1}{|J'(P_{inf}(t))|} = \frac{1}{J'(P_{inf}(t))}$, προκύπτει ότι $K_t(w_m(t), P_{disp}(t) - P_{inf}(t)) = \frac{1}{w_m(t) \cdot J'(P_{inf}(t))} \leq 1$, και άρα $h(K_t(w_m(t), P_{disp}(t) - P_{inf}(t))) = P_{disp}(t) - P_{inf}(t)$, δηλαδή η $f_u(w_m)$ παίρνει την ελάχιστη τιμή.

Περίπτωση 2: Αν $P_{disp}(t) < P_{inf}(t)$. Είναι:

$$\begin{aligned} g(w_m(t), P_{inf}(t) - P_{disp}(t)) &= \max\{P_{inf}(t) - P_{disp}(t), h(K_t(w_m(t), P_{disp}(t) - P_{inf}(t)))\} \\ &= P_{inf}(t) - P_{disp}(t) \end{aligned}$$

Γιατί εφόσον είναι:

$$\begin{aligned} K_t(w_m(t), P_{disp}(t) - P_{inf}(t)) &= \frac{1}{w_m(t) \cdot J'(P_{disp}(t) + P_{disp}(t) - P_{inf}(t))} \\ &= \frac{1}{w_m(t) \cdot J'(P_{disp}(t) - P_{disp}(t) + P_{inf}(t))} \\ &= \frac{1}{w_m(t) \cdot J'(P_{inf}(t))} \end{aligned}$$

Αφού η συνάρτηση $J(P_{PCC}(t) + P_{disp}(t))$ είναι άρτια, οπότε η παράγωγός της είναι περιττή. Όμως, επειδή η παράγωγος είναι γνησίως αύξουσα ισχύει ότι $J'(P_{inf}(t)) < 0$ και ισχύει ότι $w_m(t) \geq \frac{1}{J'(P_{inf}(t))} = -\frac{1}{J'(P_{inf}(t))}$, άρα είναι $K_t(w_m(t), P_{disp}(t) - P_{inf}(t)) \leq 1$, οπότε θα είναι $g(w_m(t), P_{inf}(t) - P_{disp}(t)) = P_{inf}(t) - P_{disp}(t)$, δηλαδή η $f_u(w_m(t))$ παίρνει την ελάχιστη τιμή. $\square$

Εφόσον είναι γνωστό ότι η η συνάρτηση $f_u(w_m(t))$ αποτελεί φράγμα για την απόκλιση $|P_{real}(t) - P_{disp}(t)|$, ενώ ο στόχος του δικτύου είναι η απόκλιση να είναι μικρότερη από την τιμή d_{max} , τότε μπορεί να επιλεγεί βάρος $w_m(t)$ τέτοιο, ώστε $f_u(w_m(t)) \leq d_{max}$. Με αυτόν τον τρόπο θα ικανοποιείται ο στόχος, εφόσον είναι εφικτός, ενώ σε διαφορετική περίπτωση θα επιλέγεται βάρος, το οποίο θα αρκεί για να δοθεί η καλύτερη λύση. Επειδή δεν πρέπει να αμελείται και το battery objective επιλέγεται το ελάχιστο βάρος που ικανοποιεί τη σχέση $f_u(w_m(t)) \leq d_{max}$.

Τελικά, τα βάρη υπολογίζονται βάσει της συνάρτησης:

$$w_m(t) = \begin{cases} 0 & , d_{max} > f_{UB} \\ \frac{1}{2|P_{inf}(t) - P_{disp}(t)|} & , d_{max} < f_{LB} \\ \frac{1}{2(d_{max} - |\Delta P|_{max})} & , \text{else} \end{cases} \quad (4.48)$$

Η πρώτη περίπτωση αντιστοιχεί στην περίπτωση όπου κάθε δυνατή δράση ικανοποιεί το στόχο του δικτύου. Αυτό ισχύει, γιατί η τιμή του f_{UB} είναι ίση με το άθροισμα της αναγκαία απόκλιση από το πλάνο διανομής λόγω των περιορισμών ($|P_{disp}(t) - P_{inf}(t)|$), της μέγιστης

απόκλισης της επιλεγείσας δράσεις από τη βέλτιστη ($\sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}|$) και της μέγιστης απόκλισης εξαιτίας της αβεβαιότητας ($|\Delta P|_{max}$). Επομένως, επιλέγεται ως βάρος η τιμή 0, ώστε η κάθε μπαταρία να ικανοποιεί την προτίμηση της.

Η δεύτερη περίπτωση χρησιμοποιείται όταν δεν είναι εφικτό να ικανοποιηθεί ο στόχος, αν η αβεβαιότητα έχει τη μέγιστη τιμή της. Επομένως, δεν υπάρχει κάποιο βάρος τέτοιο, ώστε $f_u(w_m(t)) < d_{max}$, οπότε δίνεται το ελάχιστο βάρος που δίνει το καλύτερο δυνατό φράγμα. Σύμφωνα με την ιδιότητα 4 για τη συνάρτηση $f_u(w_m(t))$ η τιμή αυτή είναι ίση με:

$$w_m(t) = \frac{1}{|J'(P_{inf}(t))|} = \frac{1}{2|P_{inf}(t) - P_{disp}(t)|}$$

Τέλος, ο υπολογισμός των βαρών γίνεται βάσει του ακόλουθου θεωρήματος, το οποίο αποδεικνύεται στο [2] και παρατίθεται εν συντομία η απόδειξή του για λόγους πληρότητας.

Θεώρημα 4.1. Έστω τα πρόβλήματα βελτιστοποίησης:

$$\min_{w_m(t) > 0} f_u(w_m(t)) \leq d_{max} \quad (4.49)$$

Και:

$$\min_{w_m(t) > 0} g(w_m, d_{max} - |\Delta P|_{max}) \leq d_{max} - |\Delta P|_{max} \quad (4.50)$$

Τότε η λύση του (4.50), αν υπάρχει, είναι και η λύση του (4.49)

Απόδειξη. Έστω w_m^* η βέλτιστη λύση του προβλήματος (4.50), και άρα ισχύει $g(w_m, d_{max} - |\Delta P|_{max}) \leq d_{max} - |\Delta P|_{max}$, τότε το w_m^* είναι εφικτή λύση του προβλήματος (4.49), καθώς:

$$\begin{aligned} f_u(w_m^*) &= \min_{d > 0} g(w_m^*, d) + |\Delta P|_{max} \\ &\leq g(w_m^*, d_{max} - |\Delta P|_{max}) + |\Delta P|_{max} \\ &\leq d_{max} \end{aligned}$$

Έστω $w'_m \in (0, w_m^*)$, τότε θα είναι $f_u(w'_m) > d_{max}$. Πράγματι, θεωρώντας μία αυθαίρετη σταθερά d , τότε:

Περίπτωση 1: Αν $d > d_{max} - |\Delta P|_{max}$, τότε:

$$\begin{aligned} g(w'_m, d) &= \max\{d, h(K_t(w'_m, d))\} \\ &\geq d \\ &> d_{max} - |\Delta P|_{max} \end{aligned}$$

Περίπτωση 2: Αν $d < d_{max} - |\Delta P|_{max}$, τότε, χρησιμοποιούνται οι ακόλουθες ιδιότητες:

- Η $K_t(w_m, d)$ είναι φθίνουσα ως προς w_m .
- Η $h(K)$ είναι αύξουσα ως προς K .

Επειδή είναι $w'_m < w_m^*$ θα είναι $K_t(w'_m, d) > K_t(w_m^*, d)$, και άρα $h(K_t(w'_m, d)) \geq h(K_t(w_m^*, d))$. Όμως w_m^* η βέλτιστη λύση του (4.50), άρα είναι $g(w_m^*, d_{max} - |\Delta P|_{max}) \leq d_{max} - |\Delta P|_{max}$.

Όμως, είναι $g(w'_m, d_{max} - |\Delta P|_{max}) = \max\{d_{max} - |\Delta P|_{max}, h(K_t(w'_m, d_{max} - |\Delta P|_{max}))\}$, άρα θα ισχύει ότι $h(K_t(w'_m, d_{max} - |\Delta P|_{max})) > d_{max} - |\Delta P|_{max} > d$. Επομένως, είναι:

$$\begin{aligned} g(w'_m, d) &= \max\{d, h(K_t(w'_m, d))\} \\ &\geq h(K_t(w'_m, d)) \\ &\geq h(K_t(w'_m, d_{max} - |\Delta P|_{max})) \\ &> d \end{aligned}$$

Άρα για κάθε $d \neq d_{max} - |\Delta P|_{max}$ ισχύει $g(w'_m, d) > d_{max} - |\Delta P|_{max}$, οπότε για κάθε $w'_m < w_m^*$ είναι $f_u(w'_m) > d_{max}$, άρα δεν είναι εφικτές λύσεις του προβλήματος (4.49), επομένως η βέλτιστη λύση είναι το βάρος w_m^* . $\square$

Για την ειδική περίπτωση της εργασίας, η βέλτιστη λύση του προβλήματος (4.50) είναι η $w_m(t) = \frac{1}{2 \cdot |d_{max} - |\Delta P|_{max}|}$. Αυτό είναι εμφανές από τη μορφή της συνάρτησης $g(w_m, d)$, για τα διάφορα βάρη, όπως παρουσιάζεται στο [Παράρτημα Β](#). Η τιμή αυτή έχει την ιδιότητα ότι ικανοποιείται οριακά η ανισότητα $K_t(w_m, d_{max} - |\Delta P|_{max})$, οπότε ουσιαστικά τα μικρότερα βάρη οδηγούν σε μεγαλύτερη τιμή για τη συνάρτηση $g(w_m, d_{max} - |\Delta P|_{max})$.

Οι υπολογισμοί της εργασίας έγιναν θεωρώντας ότι $a_p \leq 1$, όπου a_p υπενθυμίζεται ότι είναι ο συντελεστής του κόστους μπαταρίας. Στην παρούσα εργασία το a_p θεωρείται παράμετρος του προβλήματος ενισχυτικής μάθησης και επιτρέπεται να έχει και μεγαλύτερη τιμή. Στο [Παράρτημα Β](#) εξηγούνται οι προϋποθέσεις, για να εξακολουθούν οι σχέσεις και σε αυτή την περίπτωση. Η γενική ιδέα είναι ότι οι τροχιές υπολογίζονται σε σενάρια που επιλέγονται από την ίδια κατανομή με τα σενάρια της εκπαίδευσης, επομένως λειτουργούν ικανοποιητικά σε σχέση με το κόστος δικτύου. Ακόμη, το αρχικό πρόβλημα δεν περιλαμβάνει περιορισμό για την απόκλιση από τη βέλτιστη λύση, οπότε το να μην ικανοποιείται αυτός ο περιορισμός προκειμένου να είναι πιο δίκαιη η λύση δεν αποτελεί πρόβλημα. Έτσι, επιτρέπονται και μεγαλύτερες τιμές για το a_p , και θα εξεταστεί μέσω προσομοιώσεων αν λειτουργεί καλά τόσο ως προς το κόστος δικτύου όσο και ως προς τη δίκαιη χρήση μεταξύ μπαταριών.

Τα βάρη εξαρτώνται από την τιμή του $P_{inf}(t)$, επομένως βάσει του τύπου είναι διαφορετικά για κάθε σενάριο εκπαίδευσης. Όμως, εξακολουθεί να είναι συνάρτηση της συνολικής κατάστασης των πρακτόρων ενισχυτικής μάθησης, για τις δύο επιλογές που παρουσιάστηκαν στην προηγούμενη ενότητα. Γι' αυτό τον λόγο η τιμή υπολογίζεται online βάσει της εκτίμησης του φορτίου και αγνοώντας περιορισμούς σχετικά με το State of Energy.

Σχετικά με τον στόχο d_{max} πρέπει να επιλεχθεί μία τιμή που να είναι σχετικά μικρή. Οι περιορισμοί είναι να είναι μεγαλύτερος από τη μέγιστη αβεβαιότητα, δηλαδή $d_{max} > |\Delta P|_{max}$ και να είναι μικρότερος από το άνω φράγμα για την απόκλιση, δηλαδή $d_{max} \leq f_{UB}$. Ο δεύτερος περιορισμός δεν είναι αναγκαίος, όμως κρίνεται χρήσιμος, για να αποφεύγονται τα μηδενικά βάρη, όπως προϋποθέτει ο συγκεκριμένος σχεδιασμός της συνάρτησης βελτιστοποίησης. Αναφορικά με την τελική επιλογή για την τιμή του d_{max} , αυτή αφήνεται ως παράμετρος του προβλήματος ενισχυτικής μάθησης, καθώς για μικρές τιμές του μπορεί να αγνοείται το κόστος μπαταρίας, ενώ μεγάλες τιμές μπορεί να έχουν ως αποτέλεσμα να ακολουθούνται οι τροχιές χωρίς να λαμβάνονται υπόψιν οι εκτιμήσεις που προκύπτουν σε πραγματικό χρόνο.

Αναφορικά με τη μέγιστη αβεβαιότητα $|\Delta P|_{max}$ στην εργασία αυτή η τιμή προκύπτει ως η μέγιστη που παρατηρείται στα δεδομένα. Επειδή η εκτίμηση γίνεται βάσει την προηγούμενη τιμή, ως τιμή του $|\Delta P|_{max}$ ορίζεται η τιμή $\max_{d \in \{1, \dots, D\}, t \in \{k+1, \dots, T-1\}} (L_d(t) - L_d(t-1))$, όπου $L_d(t)$ το φορτίο από τα μη ελεγχόμενα στοιχεία τη χρονική στιγμή t για το d -οστό σενάριο.

Σχετικά με την προτίμηση της μπαταρίας αυτή γίνεται βάσει της συνάρτησης:

$$B^*(SoE_i(t), t) = \begin{cases} B_i^{max} & , \Delta SoE_i(t) \leq -\epsilon_i \\ -\frac{\Delta SoE_i(t) \cdot B_i^{max}}{\epsilon_i} & , |\Delta SoE_i(t)| < \epsilon_i \\ -B_i^{max} & , \Delta SoE_i(t) \geq \epsilon_i \end{cases} \quad (4.51)$$

Όπου $\epsilon_i = B_i^{max} \cdot \Delta t$ και $\Delta SoE_i(t) = SoE_i(t) - SoE_i^*(t+1)$, με $SoE_i^*(t+1)$ να συμβολίζει την τιμή για το SoE της μπαταρίας i τη χρονική στιγμή $t+1$ σύμφωνα με την τροχιά της. Η συνάρτηση αυτή δίνει την τιμή της ενεργού ισχύους, για την οποία η τιμή του SoE που προκύπτει είναι κοντινότερη δυνατή στο $SoE_i^*(t+1)$.

Η συνάρτηση ποινής $\psi_i(t)$ ορίζεται ως:

$$\psi_i(t) = \begin{cases} c_B & , SoE_i(t) < SoE_i^{min} \text{ or } SoE_i(t) > SoE_i^{max} \\ 0 & , \text{else} \end{cases} \quad (4.52)$$

Χρησιμοποιεί ώστε να ικανοποιείται ο περιορισμός για το SoE της μπαταρίας. Η τιμή c_B πρέπει να είναι θετική, ενώ αφήνεται η τελική τιμή ως παράμετρος του προβλήματος. Σχετικά με την τιμή αυτή, μικρές τιμές μπορεί να έχουν ως αποτέλεσμα να αγνοείται ο περιορισμός, ενώ μεγάλες τιμές μπορεί να μην επιτρέπουν την εξερεύνηση λύσεων που ικανοποιούν οριακά τον περιορισμό.

Αντίστοιχα, ορίζεται για τους συνδυαστικούς περιορισμούς η συνάρτηση ποινής:

$$\Lambda_i(t) = \begin{cases} c_C & , \exists j : t = t_j - 1 \text{ and } a_{ji} \neq 0 \text{ and } \sum_{k=1}^{N_B} a_{jk} SoE_k(t+1) > SoE_j^{max, t_j} \\ c_C & , \exists j : t = t_j - 1 \text{ and } a_{ji} \neq 0 \text{ and } \sum_{k=1}^{N_B} a_{jk} SoE_k(t+1) < SoE_j^{min, t_j} \\ 0 & , \text{else} \end{cases} \quad (4.53)$$

Ομοίως η τιμή c_C πρέπει να είναι θετική, ενώ η τιμή της προκύπτει βάσει των προσομοιώσεων.

Η μέθοδος αυτή σχεδιάστηκε ανεξάρτητα από τον τρόπο που έχουν υπολογιστεί οι τροχιές και μπορεί να χρησιμοποιηθεί ανεξάρτητα ποιο κριτήριο δικαιοσύνης χρησιμοποιήθηκε για τον σχεδιασμό. Αυτό που διαφέρει είναι η αιτία που αναμένεται αυτή η μέθοδος να πετύχει το στόχο της δικαιοσύνης. Έστω ότι χρησιμοποιείται για τον υπολογισμό των τροχιών η μέθοδος που παρουσιάζεται στην [υποενότητα 4.2.1](#) και χρησιμοποιεί ως hard constraint το κριτήριο δικαιοσύνης $\frac{B_i(t)}{B_i^{max}} = \frac{B_j(t)}{B_j^{max}}$. Τότε για τις βέλτιστες λύσεις που προκύπτουν από τον έλεγχο σύμφωνα με τη μέθοδο που παρουσιάστηκε στην [υποενότητα 2.4](#), ισχύει η σχέση:

$$\frac{B_i^{opt}(t)}{B_i^{max}(t)} = \frac{B_j^{opt}(t)}{B_j^{max}(t)}$$

Δηλαδή το κριτήριο δικαιοσύνης υλοποιείται με ακρίβεια. Η απόδειξη παρουσιάζεται στο [Παράρτημα Β](#). Σε περίπτωση που οι τροχιές δεν έχουν σχεδιαστεί χρησιμοποιώντας ως hard constraint το κριτήριο δίκαιης χρήσης, τότε η μέθοδος αναμένεται να λειτουργεί σχετικά

καλά, γιατί οι αποκλίσεις από αυτό θα είναι μικρές.

Από την άλλη, αν χρησιμοποιηθούν τροχιές σχεδιασμένες με μέθοδο της [υποενότητας 4.2.2](#), τότε βάσει του κριτηρίου δικαιοσύνης που χρησιμοποιήθηκε ο στόχος είναι η τιμή του SoE των μπαταριών να συγκλίνει. Έτσι, η προτίμηση δείχνει συνεχώς μία κατεύθυνση προς τη σύγκλιση, τουλάχιστον σε σχέση με τις αρχικές τιμές. Γι' αυτό και θεωρείται ότι η παραπάνω μέθοδος μπορεί να ικανοποιήσει και τα δύο κριτήρια δικαιοσύνης που χρησιμοποιούνται στην παρούσα εργασία.

4.5 Συνάρτηση Ανταμοιβής χωρίς Τροχιές

Το κριτήριο δικαιοσύνης που παρουσιάζεται [υποενότητα 2.5.2](#) θεωρεί γνωστό το αν οι μπαταρίες πρέπει να φορτίσουν ή εκφορτίσουν για μία δεδομένη χρονική στιγμή, γι' αυτό και η άμεση εφαρμογή του ως ανταμοιβή στη συνάρτηση ανταμοιβής του προβλήματος ενισχυτικής μάθησης δεν είναι πρακτική, επομένως εξετάζεται μόνο μέσω των τροχιών.

Αντίθετα, το κριτήριο της [υποενότητας 2.5.1](#) δεν προϋποθέτει κάποια γνωστή πληροφορία και έχει την ίδια μορφή σε κάθε περίπτωση. Επομένως, μπορεί να εισαχθεί άμεσα στη συνάρτηση ανταμοιβής, σύμφωνα με το πρόβλημα βελτιστοποίησης [\(4.9\)](#). Η μορφή της συνάρτησης σε αυτή την περίπτωση είναι:

$$r_i(t) = \left(\tilde{L}(t) + \sum_{j=1}^{N_B} B_j(t) - P_{disp}(t) \right)^2 + w_f \cdot \sum_{j=1}^{N_B} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2 + \psi_i(t) + \Lambda_i(t) \quad (4.54)$$

Ομοίως, η εκτίμηση για το φορτίο προκύπτει από την προηγούμενη χρονική στιγμή, δηλαδή $\tilde{L}(t) = L(t - 1)$. Ο δεύτερος όρος αποτελεί το τετραγωνικό σφάλμα της απόκλισης από την ιδανική δικαιοσύνη. Στην εργασία αυτή, ο στόχος είναι η δίκαιη χρήση μεταξύ όλων των μπαταριών, γι' αυτό και αθροίζεται το σφάλμα για όλες τις μπαταρίες. Γενικά, μπορεί η δικαιοσύνη να αφορά μόνο ένα υποσύνολο, οπότε αθροίζονται τα σφάλματα μόνο για τις αντίστοιχες μπαταρίες.

Σχετικά με το w_f , θεωρείται μία σταθερά, η οποία δίνει τη βαρύτητα στη δικαιοσύνη. Η τιμή της σταθεράς αυτής προκύπτει μέσω των προσομοιώσεων. Διαισθητικά, μεγάλη τιμή δίνει μεγαλύτερη βαρύτητα στη δίκαιη χρήση, όμως υπάρχει ο κίνδυνος στο πρόβλημα ενισχυτικής μάθησης να αγνοείται ο όρος για το κόστος δικτύου. Αντίστοιχα, για μικρές τιμές δίνεται μεγαλύτερη βαρύτητα στο grid objective, με κίνδυνο τα αποτελέσματα να μην είναι δίκαια.

Η συνάρτηση $\psi_i(t)$ αντιστοιχεί και σε αυτή την περίπτωση στη συνάρτηση ποινής για τους περιορισμούς του SoE της μπαταρίας και έχει την ίδια μορφή με αυτή που χρησιμοποιήθηκε στην προηγούμενη ενότητα:

$$\psi_i(t) = \begin{cases} c_B & , SoE_i(t) < SoE_i^{min} \text{ or } SoE_i(t) > SoE_i^{max} \\ 0 & , \text{else} \end{cases} \quad (4.55)$$

Η συνάρτηση $\Lambda_i(t)$ αφορά τη συνάρτηση ποινής για τους συνδυαστικούς περιορισμούς. Στην προηγούμενη περίπτωση η χρήση τροχιών λειτουργούσε ως proxy για τους περιορισμούς, οπότε η ποινή χρησιμοποιούταν μόνο τις χρονικές στιγμές που είχε οριστεί ο περιορισμός.

Όμως, στην ενότητα αυτή δεν χρησιμοποιούνται τροχιές, επομένως πρέπει να δίνεται ποινή και για προγενέστερες χρονικές στιγμές, ώστε να αποφευχθούν καταστάσεις, από τις οποίες δεν είναι εφικτό να ικανοποιηθεί ο περιορισμός. Έτσι, η τελική συνάρτηση που χρησιμοποιείται είναι η:

$$\Lambda_i(t) = \begin{cases} c_C & , \exists j : t = t_j - 1 \text{ and } a_{ji} \neq 0 \text{ and } \sum_{k=1}^{N_B} a_{jk} SoE_k(t+1) > SoE_j^{max, t_j} \\ c_C & , \exists j : t = t_j - 1 \text{ and } a_{ji} \neq 0 \text{ and } \sum_{k=1}^{N_B} a_{jk} SoE_k(t+1) < SoE_j^{min, t_j} \\ c_F & , \exists j : t < t_j - 1 \text{ and } a_{ji} \neq 0 \text{ and} \\ & SoE_j^{max, t_j} - \sum_{k=1}^{N_B} a_{jk} \cdot SoE_k(t+1) < \sum_{k=1}^{N_B} a_{jk} B_k^{min} \cdot \Delta T_j(t) \\ c_F & , \exists j : t < t_j - 1 \text{ and } a_{ji} \neq 0 \text{ and} \\ & SoE_j^{min, t_j} - \sum_{k=1}^{N_B} (a_{jk} \cdot SoE_k(t+1)) > \sum_{k=1}^{N_B} a_{jk} B_k^{max} \cdot \Delta T_j(t) \\ 0 & , \text{ else} \end{cases} \quad (4.56)$$

Με $\Delta T_j(t) = |t_j - t - 5| \cdot \Delta t$. Η ποινή c_C χρησιμοποιείται για τις χρονικές στιγμές που έχει οριστεί ο περιορισμός και προστίθεται, όταν δεν ικανοποιείται. Η ποινή c_F χρησιμοποιείται για να ελέγξει αν είναι εφικτό να ικανοποιηθεί ο μελλοντικός περιορισμός από την παρούσα κατάσταση. Η μορφή αυτή χρησιμοποιείται και στις εργασίες [4], [7]. Τέλος, ισχύει ότι $c_C > 0$ και $c_F \geq 0$.

4.6 Εφαρμογή του αλγορίθμου MADDPG

Στις προηγούμενες ενότητες καθορίστηκαν πλήρως όλα τα στοιχεία για την περιγραφή του προβλήματος ως ενισχυτικής μάθησης πολλαπλών πρακτόρων. Έτσι, το επόμενο βήμα είναι η εφαρμογή του αλγορίθμου στο συγκεκριμένο πρόβλημα. Στην ενότητα αυτή θα δοθεί μία αφαιρετική περιγραφή της προσέγγισης, υπό την έννοια ότι ο αλγόριθμος λειτουργεί ανεξάρτητα από την αναπαράσταση της κατάστασης και τη συνάρτηση ανταμοιβής, από αυτές που καθορίστηκαν προηγουμένως, που επιλέγονται στην εκάστοτε εφαρμογή.

Αρχικά, για κάθε μπαταρία αντιστοιχεί ένας πράκτορας, ο οποίος αποτελείται από έναν actor, ο οποίος αναπαριστά την πολιτική για την επιλογή δράσης, και έναν critic, ο οποίος χρησιμοποιείται μόνο κατά την εκπαίδευση για να αξιολογήσει την πολιτική. Ακόμη, για λόγους σταθερότητας στην εκπαίδευση χρησιμοποιείται ο target actor και ο target critic, οι οποίοι είναι αντίγραφα των actor και critic αντίστοιχα, που μεταβάλλονται με σημαντικά χαμηλότερο ρυθμό.

Ο actor δέχεται ως είσοδο την κατάσταση της μπαταρίας που αντιστοιχεί στον agent και βάσει αυτής αποφασίζει τη δράση της. Ο actor χρησιμοποιείται για την επιλογή δράσεων και κατά την εκτέλεση, επομένως δεν έχει πρόσβαση σε πληροφορία των υπόλοιπων μπαταριών.

Από την άλλη ο critic χρησιμοποιείται μόνο κατά την εκπαίδευση. Γι' αυτό και έχει πρόσβαση στην κατάσταση όλων των πρακτόρων. Επειδή υπάρχει κοινή πληροφορία μεταξύ των πρακτόρων, στην εργασία αυτή χρησιμοποιείται η ελάχιστη διάσταση για την περιγραφή της συνολικής κατάστασης. Δηλαδή αν η κατάσταση κάθε πράκτορα περιγράφεται ως $s_i(t) = (SoE_i(t), \tilde{L}(t), t)$, τότε η συνολική κατάσταση είναι $x_s(t) = (SoE_1(t), \dots, SoE_{N_B}(t), \tilde{L}(t), t)$. Αντίστοιχα, αν χρησιμοποιείται η κατάσταση $s'_i(t) = (SoE_i(t), P(t-1), B_i(t-1), t)$, τότε συνολική κατάσταση είναι η πλειάδα $x'_s = (SoE_1(t), \dots, SoE_{N_B}(t), B_1(t-1), \dots, B_{N_B}(t-1), P(t-1), t)$.

Αυτή η αναπαράσταση μπορεί να μειώσει το χρόνο εκπαίδευσης, αν το πλήθος μπαταριών είναι μεγάλο, ενώ περιορίζεται το πλήθος των παραμέτρων προς εκμάθηση.

Σχετικά με τα δεδομένα εκπαίδευσης, χρησιμοποιείται ένα σύνολο από πιθανά φορτία για το χρονικό διάστημα του ελέγχου. Το σύνολο αυτό πρέπει να αποτελείται από αρκετά πιθανά φορτία, ώστε να μπορεί να μάθει ο πράκτορας για παρόμοια φορτία εντός ενός επιθυμητού εύρους. Η χρήση προκαθορισμένου συνόλου σεναρίων επιτρέπει ο υπολογισμός των βαρών που έχουν οριστεί για τις συναρτήσεις ανταμοιβής που χρησιμοποιούν τροχιές να γίνεται πριν την εκτέλεση των επεισοδίων, καθώς είναι συνάρτηση μόνο του φορτίου και των δύο παραμέτρων που είναι σταθερές σε κάθε περίπτωση. Με αυτό τον τρόπο αποφεύγονται άσκοποι επανυπολογισμοί.

Στην εργασία αυτή, το πλάνο διανομής, οι τροχιές των μπαταριών (εφόσον χρησιμοποιούνται) και οι περιορισμοί είναι συγκεκριμένα και τα ίδια για κάθε επεισόδιο. Σε κάθε εκτέλεση του αλγορίθμου εκπαίδευσης εκτελούνται προκαθορισμένος αριθμός επεισοδίων, ο οποίος θα αναφέρεται ως *NUM_EPISODES*. Όμως, αλγόριθμοι όπως ο MADDPG δεν εξασφαλίζουν ότι η πολιτική βελτιώνεται σε κάθε βήμα εκπαίδευσης, οπότε το μοντέλο που επιλέγεται τελικά δεν είναι αυτό που έχει μάθει τη στιγμή του τερματισμού του αλγορίθμου, αλλά αυτό της χρονικής στιγμής που η μέση συνολική ανταμοιβή για κάποια (έστω *n_recent_episodes*) προηγούμενα επεισόδια είναι η μέγιστη.

Τέλος, οι χρονικές στιγμές που έχει οριστεί συνδυαστικός περιορισμός είναι σημαντικές και έχουν ιδιαίτερη βαρύτητα σε σχέση με τις υπόλοιπες. Αν δεν δοθούν αρκετά δεδομένα για αυτές είναι πιθανό να μην μαθαίνονται και άρα να μην ικανοποιούνται κατά την εκτέλεση. Γι' αυτό τον λόγο ακολουθείται η προσέγγιση των [4], [7], όπου χρησιμοποιούνται δύο Replay Buffers. Στον πρώτο, ο οποίος ονομάζεται Terminal Replay Buffer αποθηκεύονται μεταβάσεις προς τις χρονικές στιγμές που αφορούν περιορισμούς, ενώ στον δεύτερο όλες οι υπόλοιπες. Με αυτό τον τρόπο διασφαλίζεται το ποσοστό των δεδομένων εκπαίδευσης που αφορά τους περιορισμούς.

Ακολουθεί η περιγραφή της μεθόδου εκπαίδευσης σε ψευδοκώδικα.

ΑΛΓΟΡΙΘΜΟΣ 4.2: Αλγόριθμος Εκπαίδευσης

```

for all agents do
    Initialize critic  $Q_i(x_s, B; \phi_i)$  and actor  $\mu_i(s_i; \partial_i)$  networks with weights  $\phi_i$  and  $\partial_i$ 
    Initialize target critic  $Q'_i(x_s, B; \phi'_i)$  and target actor  $\mu'_i(s_i; \partial'_i)$  with weights  $\phi_i$  and  $\partial_i$ 
end for
Initialize the Replay Buffer  $\mathcal{D}$  and the Terminal Replay Buffer  $\mathcal{D}_T$  as empty
Define  $P_{disp}$ , the set of loads  $L^{real} \leftarrow \{L_1^{real}, \dots, L_D^{real}\}$  and the trajectory for each battery
(if needed)
Define score_history as an empty list
 $best\_avg\_score \leftarrow -\infty$ 
for ( $n \leftarrow 0 \cdot n < NUM\_EPISODES \cdot ++ n$ ) do
     $score \leftarrow 0$ 
    Choose uniformly the scenario  $d$  from  $\{1, 2, \dots, |L^{real}|\}$ 
    for all agents do
        Receive initial observation  $s_i(k)$  and total state  $S \leftarrow x_S(k)$ 
    end for
    for ( $t \leftarrow k \cdot t < T \cdot ++ t$ ) do
        if  $t = k$  then
             $\tilde{L}(t) \leftarrow L_d^{real}(t)$ 
        else
             $\tilde{L}(t) \leftarrow L_d^{real}(t - 1)$ 
        end if
        for all agents do
            Select action  $b_i = \mu_i(s_i(t); \partial_i)$ 
            Choose the noise  $g_i$ 
             $b_i \leftarrow b_i + g_i$ 
             $B_i(t) \leftarrow \text{projection of } b_i \text{ to } [B_i^{min}, B_i^{max}]$ 
            Execute action  $B_i(t)$ 
            Observe new state  $s_i(t + 1)$ 
        end for
        for all agents do
            Compute reward  $r_i(t)$ 
        end for
        Compute  $x'_S(t)$  based on  $s'_i(t)$ 
         $S' \leftarrow x'_S(t)$ ,  $B \leftarrow (B_1(t), \dots, B_{N_B}(t))$ ,  $R \leftarrow (r_1(t), \dots, r_{N_B}(t))$ 
        if  $t \in \mathcal{T}$  then ▷  $\mathcal{T}$  the set of time steps with a combinatorial constraint
            Store  $(S, B, S', R)$  in Terminal Replay Buffer  $\mathcal{D}_T$ 
        else
            Store  $(S, B, S', R)$  in Replay Buffer  $\mathcal{D}$ 
        end if
        Run MADDPG training
         $score \leftarrow score + \sum_{i=1}^{N_B} r_i$ 
         $S \leftarrow S'$ 
    end for
     $score\_history.append(score)$ 
    if  $avg(score\_history[-n\_recent\_episodes :]) > best\_avg\_score$  then
        Save the weights
         $best\_avg\_score \leftarrow avg(score\_history[-n\_recent\_episodes :])$ 
    end if
end for

```

Σχετικά με την εφαρμογή του αλγορίθμου στο στάδιο της εκτέλεσης, το μόνο που χρειάζεται να είναι γνωστό είναι τα βάρη του κύριου actor κάθε πράκτορα. Βάσει αυτών ο κάθε πράκτορας λαμβάνει απόφαση για την ενεργό ισχύ του, χωρίς να λαμβάνει πληροφορία για τους υπόλοιπους. Η μόνη πληροφορία που απαιτείται είναι για το φορτίο ή τη συνολική ισχύ στο PCC την αμέσως προηγούμενη χρονική στιγμή, ανάλογα με την αναπαράσταση που χρησιμοποιείται.

Ο αλγόριθμος θα αξιολογηθεί τόσο ως προς το grid objective όσο και ως προς τη δίκαιη χρήση που εξασφαλίζει. Σχετικά με το κόστος του δικτύου, αν κατά την εκτέλεση παρατηρείται φορτίο $L^{real}(t)$, τότε το κόστος είναι:

$$cost = \sum_{t=k}^{T-1} (L^{real}(t) + \sum_{i=1}^{N_B} B_i(t) - P_{disp}(t))^2 \quad (4.57)$$

Όσο μικρότερο είναι το κόστος τόσο καλύτερη είναι η πολιτική που έχει μάθει. Το κόστος αυτό πρέπει να συγκριθεί με το ελάχιστο δυνατό, το οποίο μπορεί να υπολογιστεί λύνοντας offline το αντίστοιχο πρόβλημα βελτιστοποίησης θεωρώντας γνωστό εξ αρχής όλο το φορτίο. Επίσης, αν χρησιμοποιούνται τροχιές, ένας επιτυχημένος αλγόριθμος θα έχει μικρότερο κόστος, τουλάχιστον για κάποια φορτία, από το κόστος που θα παρατηρούνταν, αν κάθε μπαταρία ακολουθούσε την τροχιά της.

Αναφορικά με το κριτήριο δικαιοσύνης που παρουσιάστηκε στην [υποενότητα 2.5.1](#), ο αλγόριθμος θα αξιολογηθεί ως προς το μέσο τετραγωνικό σφάλμα για την απόκλιση από τη δίκαιη χρήση. Θεωρώντας ότι ο στόχος είναι η δικαιοσύνη μεταξύ όλων των μπαταριών, τότε η κατάλληλη μετρική αξιολόγησης είναι η:

$$fairness_metric_1 = \frac{1}{T-k} \cdot \frac{1}{\binom{N_B}{2}} \cdot \frac{1}{4} \cdot \sum_{t=k}^{T-1} \sum_{i=1}^{N_B} \sum_{j=1}^{N_B} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2 \quad (4.58)$$

Η τιμή 0 δηλώνει την απολύτως δίκαιη χρήση, ενώ τιμές κοντά στο 1 δηλώνουν ότι η χρήση δεν είναι δίκαιη.

Αντίστοιχα, για το κριτήριο της υποενότητας [υποενότητα 2.5.2](#) μπορεί να οριστεί το ανάλογο κριτήριο ως:

$$fairness_metric = \frac{1}{T-k} \cdot \left[\sum_{t \in T_1} \sum_{i=1}^{N_B} \sum_{j=1}^{N_B} \left(B_i(t) \cdot (SoE_j(t) - SoE_j^{min}) - B_j(t) \cdot (SoE_i(t) - SoE_i^{max}) \right)^2 + \sum_{t \in T_2} \sum_{i=1}^{N_B} \sum_{j=1}^{N_B} \left(B_i(t) \cdot (SoE_j^{max} - SoE_j(t)) - B_j(t) \cdot (SoE_i^{min} - SoE_i(t)) \right)^2 \right] \quad (4.59)$$

Όπου T_1 το σύνολο των χρονικών στιγμών με $\sum_{i=1}^{N_B} B_i(t) < 0$ και αντίστοιχα T_2 το σύνολο των χρονικών στιγμών με $\sum_{i=1}^{N_B} B_i(t) > 0$. Το κριτήριο αυτό μπορεί σε κάποιες περιπτώσεις να είναι biased, όπως αν γίνεται εκφόρτιση κάθε χρονική στιγμή λύσεις με μικρά SoE ευνοούνται χωρίς απαραίτητα αυτό να σημαίνει δικαιοσύνη. Ένα απλούστερο κριτήριο είναι να ελέγχεται αν έχουν συγκλίνει οι τιμές των SoE των μπαταριών. Μία μετρική είναι η

ακόλουθη:

$$fairness_metric_2 = \frac{1}{\binom{N_B}{2}} \sum_{i=1}^{N_B} \sum_{j \neq i} \left| \frac{SoE_i(T) - SoE_j(T)}{SoE_i(k) - SoE_j(k)} \right| \quad (4.60)$$

Στόχος είναι η παραπάνω μετρική να έχει τιμή μικρότερη από 1. Για τις περιπτώσεις που εξετάζονται θεωρείται ότι όλες οι μπαταρίες εκκινούν από διαφορετικό SoE, οπότε ορίζεται η μετρική αυτή.

Ο αλγόριθμος που χρησιμοποιείται για την αξιολόγηση των μοντέλων σε μορφή ψευδο-κώδικα είναι ο εξής:

ΑΛΓΟΡΙΘΜΟΣ 4.3: Αλγόριθμος Αξιολόγησης

```

for all agents do
  Load the learned weights  $\partial_i$  of the actor  $\mu_i(s_i; \partial_i)$ 
end for
Define  $P_{disp}$ , the load  $L^{real}$  and the trajectory for each battery (if needed)
for all agents do
  Receive initial observation  $s_i(k)$ 
end for
 $cost \leftarrow 0$ 
 $fairness\_metric_1 \leftarrow 0$ 
 $fairness\_metric_2 \leftarrow 0$ 
for ( $t \leftarrow k$   $t < T$   $++ t$ ) do
  if  $t = k$  then
     $\tilde{L}(t) \leftarrow$  initial guess if prior data are available, else 0
  else
     $\tilde{L}(t) \leftarrow L^{real}(t - 1)$ 
  end if
  for all agents do
    Select action  $B_i(t) = \mu_i(s_i(t); \partial_i)$ 
    Execute action  $B_i(t)$ 
    Observe new state  $s_i(t + 1)$ 
  end for
  for all agents  $i$  do
    for all agents  $j \neq i$  do
       $fairness\_metric_1 \leftarrow fairness\_metric_1 + \left( \frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2$ 
    end for
  end for
   $cost \leftarrow cost + (\sum_{i=1}^{N_B} B_i(t) + L^{real}(t) - P_{disp}(t))^2$ 
end for
 $fairness\_metric_1 \leftarrow \frac{fairness\_metric_1}{(T-k) \cdot \binom{N_B}{2} \cdot 4}$ 
for all agents  $i$  do
  for all agents  $j \neq i$  do
     $fairness\_metric_2 \leftarrow fairness\_metric_2 + abs \left( \frac{SoE_i(T) - SoE_j(T)}{SoE_i(k) - SoE_j(k)} \right)$ 
  end for
end for
 $fairness\_metric_2 \leftarrow \frac{fairness\_metric_2}{\binom{N_B}{2}}$ 

```

Κεφάλαιο 5

Υλοποίηση

Στο κεφάλαιο αυτό περιγράφεται η υλοποίηση που χρησιμοποιήθηκε, για την εκτέλεση υπολογιστικών προσομοιώσεων, με στόχο την αξιολόγηση των μεθόδων που έχουν προταθεί προς τη δίκαιη χρήση των μπαταριών. Αρχικά, αναλύονται οι βιβλιοθήκες της γλώσσας python που χρησιμοποιήθηκαν και παρουσιάζονται αναλυτικά οι περιπτώσεις που εξετάζονται με σκοπό την αξιολόγηση των μεθόδων. Έπειτα, γίνονται συγκεκριμένοι οι συνδυασμοί αναπαράστασης της κατάσταση και της συνάρτησης που αξιολογούνται, βάσει των προτάσεων του κεφαλαίου 4. Στη συνέχεια, γίνονται γνωστές οι παράμετροι του προβλήματος ενισχυτικής μάθησης, δίνοντας μία διαισθητική κατεύθυνση για την επιλογή, όπου αυτό είναι εφικτό. Τέλος, δίνονται τα αποτελέσματα που προέκυψαν μέσω αυτών των προσομοιώσεων και αξιολογούνται ως προς τους στόχους τους.

5.1 Λεπτομέρειες Υλοποίησης

Η υλοποίηση πραγματοποιήθηκε στη γλώσσα Python3. Οι κύριες βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι:

- **NumPy**[33]: Για τη διαχείριση πινάκων.
- **Matplotlib**[34]: Για τη σχεδίαση γραφικών παραστάσεων.
- **TensorFlow**[35]: Για την ανάπτυξη του αλγορίθμου ενισχυτικής μάθησης και την υλοποίηση των νευρωνικών δικτύων.
- **Gymnasium**[36]: Για τον σχεδιασμό του περιβάλλοντος του προβλήματος ενισχυτικής μάθησης.

Επίσης, για τον σχεδιασμό των τροχιών χρησιμοποιήθηκε η βιβλιοθήκη SciPy. Η βιβλιοθήκη αυτή περιλαμβάνει υλοποιήσεις μεθόδων βελτιστοποίησης. Εφόσον τα προβλήματα υπολογισμού τροχιών έχουν οριστεί ως προβλήματα κυρτής βελτιστοποίησης, μπορούν να χρησιμοποιηθούν μέθοδοι τοπικής βελτιστοποίησης που παρέχονται από τη βιβλιοθήκη. Συγκεκριμένα, τα προβλήματα επιλύονται με χρήση της μεθόδου Sequential Least Squares Programming (SLSQP), η οποία είναι η προτεινόμενη για προβλήματα βελτιστοποίησης βαθμωτών συναρτήσεων πολλαπλών μεταβλητών με περιορισμούς. Η μέθοδος αυτή είναι ιδανική για προβλήματα όπου η συνάρτηση βελτιστοποίησης και οι συναρτήσεις των περιορισμών είναι δύο φορές παραγωγίσιμοι, ενώ λειτουργεί αποδοτικά για μεγάλο αριθμό μεταβλητών και

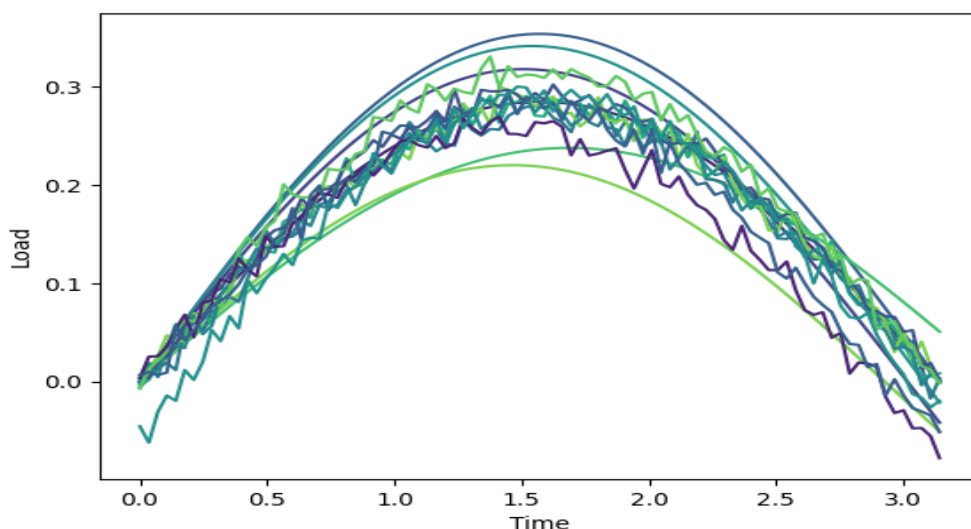
οποιοδήποτε συνδυασμό από περιορισμούς ισοτήτων και ανισοτήτων [37], οπότε καθίσταται ως κατάλληλη για την περίπτωση.

5.2 Περιπτώσεις Χρήσης

Η παρούσα εργασία εξετάζει το επίπεδο ελέγχου σε πραγματικό χρόνο. Έτσι, οι εξεταζόμενες περιπτώσεις αφορούν ένα χρονικό παράθυρο ελέγχου των 15 λεπτών. Στην εργασία αυτή, το χρονικό διάστημα αυτό διαιρείται σε 90 μικρότερα διαστήματα, με το κάθε ένα να έχει διάρκεια $\Delta t = 10\text{sec}$. Για να είναι εφικτός ο έλεγχος σε πραγματικό χρόνο πρέπει το μοντέλο που εκπαιδεύτηκε να υπολογίζει μέσα σε αυτό το χρονικό διάστημα την τιμή της ισχύος για το επόμενο χρονικό διάστημα των 10 δευτερολέπτων. Στον χρόνο αυτό συμπεριλαμβάνεται και ο χρόνος να ενημερωθεί ο πράκτορας για την κοινή πληροφορία από το PCC.

5.2.1 Σενάρια Εκπαίδευσης

Τα σενάρια που χρησιμοποιούνται για την εκπαίδευση, αλλά και για τον υπολογισμό των τροχιών, έχουν ημιτονοειδή μορφή. Για το πλάτος οι τιμές επιλέγονται τυχαία από 0.21 έως 0.39 ανά μονάδα. Αντίστοιχα, η περίοδος κυμαίνεται μεταξύ 27.2min και 33.3min και επιλέγεται επίσης ομοιόμορφα. Ακόμη, η αρχική φάση του ημιτόνου παίρνει τιμές στο διάστημα $[-\frac{\pi}{12}, \frac{\pi}{12}]$. Τέλος, σε κάποια από τα σενάρια προστίθεται σε κάθε χρονική στιγμή τυχαίος θόρυβος. Στην ακόλουθη γραφική παράσταση φαίνονται κάποιο πιθανά σενάρια που προκύπτουν από την κατανομή που περιγράφηκε.



Σχήμα 5.1: Πιθανά Σενάρια Εκπαίδευσης

5.2.2 Περίπτωση 1: 2 Μπαταρίες και Βέλτιστο Πλάνο Διανομής

Η πρώτη περίπτωση που εξετάζεται εστιάζει σε δίκτυο που περιλαμβάνει 2 μπαταρίες και μοναδικό σημείο PCC. Για την περίπτωση αυτή το πλάνο διανομής και οι τροχιές που πρέπει να ακολουθούν οι μπαταρίες υπολογίζονται, σύμφωνα με τη [μέθοδο της ενότητας 2.3](#). Συγκεκριμένα, επιλύεται το παρακάτω πρόβλημα βελτιστοποίησης:

$$\begin{aligned}
 & \min_{B_1(t), B_2(t), P_{disp}(t)} \sum_{t=0}^{T-1} \left\{ \frac{1}{D} \sum_{d=1}^D [P_{PCC,d}(t) - P_{disp}(t)]^2 \right\} \\
 & \text{subject to } P_{PCC,d}(t) = \sum_{i=1}^2 B_i(t) + L_d(t) \\
 & \quad SoE_i(t+1) = SoE_i(t) + B_i(t)\Delta t \\
 & \quad B_i^{min} \leq B_i(t) \leq B_i^{max} \\
 & \quad SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \\
 & \quad SoE_A^{min,T} \leq SoE_1(T) + SoE_2(T) \leq SoE_A^{max,T} \\
 & \quad \text{for } i \in \{1, 2\}, t \in \{0, \dots, T-1\}
 \end{aligned}$$

Οι τιμές για τα χαρακτηριστικά των στοιχείων του δικτύου αναφέρονται στο [Παράρτημα Γ](#), με τις μπαταρίες 1 και 2 να χρησιμοποιούνται για την παρούσα περίπτωση. Επειδή παρατηρήθηκε ότι για το πρόβλημα που προέκυπτε ο τελικός περιορισμός ικανοποιούνταν πάντα με ευκολία, ακόμα και για μεθόδους που δεν λάμβαναν υπόψιν τον τελικό περιορισμό, αυτός αντικαταστήθηκε από τον περιορισμό:

$$SoE_1^{tr}(T) + SoE_2^{tr}(T) \leq SoE_1(T) + SoE_2(T) \leq SoE_A^{max,T}$$

Όπου $SoE_i^{tr}(T)$ η τιμή για την κατάσταση της ενέργειας της μπαταρίας i τη χρονική στιγμή T , σύμφωνα με την τροχιά που υπολογίστηκε.

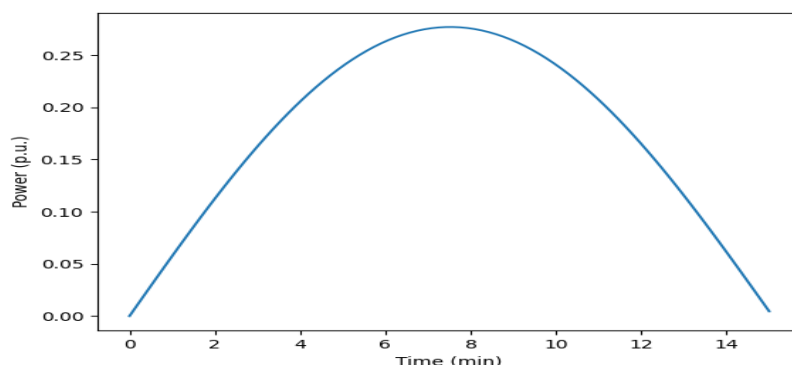
Σημειώνεται επίσης ότι οι τροχιές που παράχθηκαν σε αυτό το βήμα δεν είναι απαραίτητα δίκαιες, οπότε πιθανώς να μην είναι αυτές που θα χρησιμοποιηθούν κατά την εκπαίδευση.

Μέσω αυτού του τρόπου εξετάζεται μία περίπτωση όπου το πλάνο διανομής είναι κοντά στα αναμενόμενα φορτία, οπότε οι μπαταρίες δεν αναμένεται να λειτουργούν κοντά στα όρια ισχύος τους. Το πλάνο διανομής που εξετάστηκε στην εργασία αυτή και παράχθηκε με τη προηγούμενη μέθοδο απεικονίζεται στην ακόλουθη γραφική παράσταση:

5.2.3 Περίπτωση 2: Αυτόνομη λειτουργία δικτύου 2 μπαταριών

Στη δεύτερη περίπτωση που εξετάζεται το πλάνο διανομής θεωρείται ίσο με μηδέν για κάθε χρονική στιγμή. Αυτή η περίπτωση αντιστοιχεί σε ανεξάρτητη λειτουργία του δικτύου, δηλαδή οι ανάγκες των καταναλωτών του δικτύου καλύπτονται πλήρως από την παραγωγή και τις μπαταρίες του ίδιου δικτύου. Στην παρούσα περίπτωση στο δίκτυο συμμετέχουν 2 μπαταρίες.

Το δίκτυο που εξετάζεται είναι το ίδιο με αυτό της προηγούμενης περίπτωσης, δηλαδή



Σχήμα 5.2: Πλάνο Διανομής

ισχύουν τα ίδια χαρακτηριστικά. Σχετικά με τον τελικό περιορισμό, αυτός έχει τη μορφή:

$$SoE_B^{min,T} \leq SoE_1(T) + SoE_2(T) \leq SoE_B^{max,T}$$

Δηλαδή και σε αυτή την περίπτωση υπάρχει τελικός περιορισμός που αφορά και τις 2 μπαταρίες.

Με αυτή την περίπτωση εξετάζεται πως λειτουργούν οι προτεινόμενες μέθοδοι σε μία περίπτωση όπου ο στόχος δικτύου έρχεται σε αντίθεση με τους περιορισμούς. Έτσι, εξετάζεται μία δυσκολότερη περίπτωση όσον αφορά την ικανοποίηση των τελικών περιορισμών. Επίσης, στην περίπτωση αυτή αναμένεται οι μπαταρίες να λειτουργούν σε τιμές ισχύος κοντά στα όρια τους.

5.2.4 Περίπτωση 3: Αυτόνομη λειτουργία δικτύου 4 μπαταριών

Η τελευταία περίπτωση που θα εξεταστεί αφορά δίκτυο το οποίο αποτελείται από 4 μπαταρίες. Επίσης, το πλάνο διανομής ορίζεται ίσο με το μηδέν για όλο το χρονικό διάστημα ελέγχου, με στόχο την αυτόνομη λειτουργία του δικτύου.

Σχετικά με τον τελικό περιορισμό, αυτός αφορά και τις 4 μπαταρίες του δικτύου. Δηλαδή η μορφή του είναι η ακόλουθη:

$$SoE_C^{min,T} \leq SoE_1(T) + SoE_2(T) + SoE_3(T) + SoE_4(T) \leq SoE_C^{max,T}$$

Αυτή η περίπτωση εξετάζει την κλιμακωσιμότητα των μεθόδων, αλλά και τη δίκαιη χρήση σε περιβάλλον πολλαπλών μπαταριών. Λόγω του μεγαλύτερου πλήθους μπαταριών αυξάνονται οι δυνατότητες του δικτύου, οπότε ενώ οι τροχιές δεν είναι οι βέλτιστες για την περίπτωση χωρίς τους περιορισμούς, ο στόχος να ακολουθείται το πλάνο διανομής είναι πιο εύκολος σε σχέση με την περίπτωση 2.

5.3 Μέθοδοι Προς Αξιολόγηση

Οι μέθοδοι που εξετάζονται στην παρούσα εργασία αφορούν τους συνδυασμούς των αναπαράστασεων της κατάστασης και των συναρτήσεων ανταμοιβής που σχεδιάστηκαν στο

κεφάλαιο 4 με στόχο τη δίκαιη χρήση. Σχετικά με τις συναρτήσεις ανταμοιβής που χρησιμοποιούν τροχιές εξετάζονται δύο μέθοδοι, μία για κάθε κριτήριο δικαιοσύνης που λαμβάνεται υπόψιν κατά τον σχεδιασμό των τροχιών. Επίσης, εξετάζεται η χρήση ανταμοιβής χωρίς κανένα κριτήριο δικαιοσύνης, για να αποτελέσει αναφορά για τη σύγκριση των υπολοίπων μεθόδων και αν καταφέρνουν να πετύχουν τον επιθυμητό στόχο.

Για κάθε μέθοδο ορίζεται αναλυτικά η συνολική κατάσταση και οι παρατηρήσεις κάθε πράκτορα και δίνεται η συνάρτηση ανταμοιβής. Για τη συνάρτηση ανταμοιβής επισημαίνονται οι τιμές που αφήνονται ως παράμετροι και πρέπει να βελτιστοποιηθούν μέσω των προσομοιώσεων.

Μέθοδος 1

Η πρώτη μέθοδος που εξετάζεται χρησιμοποιεί την κατάσταση:

$$s^1(t) = (SoE_1(t), \dots, SoE_{N_B}(t), \tilde{L}(t), t)$$

Ενώ η συνάρτηση ανταμοιβής ορίζεται ως:

$$r_i^1(t) = \left(\tilde{L}(t) + \sum_{j=1}^{N_B} B_j(t) - P_{disp}(t) \right)^2 + \psi_i(t) + \Lambda_i(t)$$

Όπου οι συναρτήσεις $\psi_i(t)$ και $\Lambda_i(t)$ υπολογίζονται σύμφωνα με τις σχέσεις (4.55) και (4.56) αντίστοιχα. Ο κάθε πράκτορας έχει πρόσβαση κατά την εκτέλεση μόνο στην πληροφορία $o_i^1(t) = (SoE_i(t), \tilde{L}(t), t)$. Η μέθοδος αυτή συμπεριλαμβάνει στην ανταμοιβή μόνο το κόστος για το δίκτυο και τους περιορισμούς, οπότε δεν αναμένεται να δώσει δίκαια αποτελέσματα. Όμως, εξετάζεται ως μέθοδος αναφοράς, ώστε να εξεταστεί αν οι συναρτήσεις που σχεδιάστηκαν βελτιώνουν τη δίκαιη χρήση και αν αυτό γίνεται εις βάρος του αρχικού στόχου. Τέτοιες μέθοδοι που δεν χρησιμοποιούν κάποιο κριτήριο δικαιοσύνης θα αναφέρονται στη συνέχεια και ως βασικές.

Μέθοδος 2

Η δεύτερη μέθοδος που εξετάζεται είναι η χρήση της δεύτερης προτεινόμενης αναπαράστασης της κατάστασης για την ίδια ανταμοιβή. Δηλαδή η συνολική κατάσταση είναι η:

$$s^2(t) = (SoE_1(t), B_1(t-1), \dots, SoE_{N_B}(t), B_{N_B}(t-1), P_{PCC}(t-1), t)$$

Κάθε πράκτορας λαμβάνει παρατήρηση $o_i^2(t) = (SoE_i(t), B_i(t-1), P_{PCC}(t-1), t)$, ενώ η ανταμοιβή είναι $r_i^2(t) = r_i^1(t)$. Επίσης, αυτή η μέθοδος χρησιμοποιείται ως αναφορά και δεν λαμβάνει υπόψιν τα κριτήρια δικαιοσύνης.

Μέθοδος 3

Η τρίτη μέθοδος που χρησιμοποιείται χρησιμοποιεί τροχιές με στόχο τη δίκαιη χρήση μεταξύ των μπαταριών. Η μέθοδος αυτή στοχεύει στην επίτευξη της δικαιοσύνης σύμφωνα με το κριτήριο της **υποενότητας 2.5.1**. Για τον στόχο αυτό χρησιμοποιούνται οι δικαιότερες δυνατές τροχιές για την κατάσταση της ενέργειας των μπαταριών. Έτσι, οι τροχιές υπολογίζονται επιλύοντας το πρόβλημα ελαχιστοποίησης της συνάρτησης (4.12) υπό τους περιορισμούς (2.2), (2.6), (4.3) - (4.6) και (4.13).

Η τροχιά αυτή εισέρχεται στο πρόβλημα ενισχυτικής μάθησης μέσω της συνάρτησης

ανταμοιβής. Η συνάρτηση αυτή θα έχει τη μορφή:

$$r_i^3(t) = \left(\tilde{L}(t) + \sum_{j=1}^{N_B} B_j(t) - P_{disp}(t) \right)^2 + \frac{a_p}{w_m(t)} \frac{(B_i(t) - B_i^*(SoE_i(t), t))^2}{4B_i^{max}} + \psi_i(t) + \Lambda_i(t)$$

Με τις συναρτήσεις $\psi_i(t), \Lambda_i(t)$ να ορίζονται από τις σχέσεις (4.52) και (4.53) αντίστοιχα. Υπενθυμίζεται ότι η τιμή a_p αφήνεται ως παράμετρος του προβλήματος ενισχυτικής μάθησης. Επίσης, παράμετρος του προβλήματος είναι η τιμή του στόχου για το κόστος δικτύου κάθε χρονική στιγμή d_{max} .

Η μέθοδος 3 χρησιμοποιεί για την αναπαράσταση της κατάστασης τη πλειάδα:

$$s^3(t) = (SoE_1(t), \dots, SoE_{N_B}(t), \tilde{L}(t), t)$$

Ενώ κάθε πράκτορας λαμβάνει παρατήρηση $o_i^3(t) = (SoE_i(t), \tilde{L}(t), t)$.

Μέθοδος 4

Η μέθοδος 4 αφορά την εφαρμογή της μεθόδου 3 χρησιμοποιώντας την εναλλακτική αναπαράσταση για την κατάσταση. Δηλαδή οι τροχιές για το SoE υπολογίζεται ομοίως και η συνάρτηση ανταμοιβής για τη μπαταρία i είναι η $r_i^4(t) = r_i^3(t)$. Ενώ η κατάσταση ορίζεται ως $s^4(t) = (SoE_1(t), B_1(t-1), \dots, SoE_{N_B}(t), B_{N_B}(t-1), P_{PCC}(t-1), t)$, με τον κάθε πράκτορα να λαμβάνει παρατήρηση $o_i^4(t) = (SoE_i(t), B_i(t-1), P_{PCC}(t-1), t)$.

Μέθοδος 5

Η πέμπτη μέθοδος που εξετάζεται προσπαθεί να πετύχει το [κριτήριο δικαιοσύνης 2.5.2](#). Η λογική που ακολουθεί είναι αυτή της μεθόδου 3 με μοναδική διαφοροποίηση στον τρόπο υπολογισμού των τροχιών. Συγκεκριμένα, οι τροχιές υπολογίζονται μέσω της μεθόδου που παρουσιάστηκε στην [ενότητα 4.2.2](#) με χρήση του [αλγορίθμου 4.1](#).

Στη μέθοδο αυτή επιλέγεται η αναπαράσταση $s^5(t) = (SoE_1(t), \dots, SoE_{N_B}(t), \tilde{L}(t), t)$ και, όπως και στις προηγούμενες περιπτώσεις, ο κάθε πράκτορας λαμβάνει την παρατήρηση $o_i^5(t) = (SoE_i(t), \tilde{L}(t), t)$. Σχετικά με τη συνάρτηση ανταμοιβής ισχύει ότι $r_i^5(t) = r_i^3(t)$ για κάθε μπαταρία i .

Μέθοδος 6

Αποτελεί την εφαρμογή της μεθόδου 6 για την εναλλακτική αναπαράσταση της κατάστασης, δηλαδή αν είναι $s^6(t) = (SoE_1(t), B_1(t-1), \dots, SoE_{N_B}(t), B_{N_B}(t-1), P_{PCC}(t-1), t)$ με τον πράκτορα i να έχει γνώση της πληροφορίας $o_i^6(t) = (SoE_i(t), B_i(t-1), P_{PCC}(t-1), t)$. Χρησιμοποιείται η συνάρτηση ανταμοιβής της προηγούμενης μεθόδου, δηλαδή $r_i^6(t) = r_i^5(t)$ για κάθε μπαταρία i .

Μέθοδος 7

Η μέθοδος αυτή χρησιμοποιεί τη συνάρτηση ανταμοιβής που παρουσιάζεται αναλυτικά στην [ενότητα 4.5](#). Συγκεκριμένα, κάθε μπαταρία λαμβάνει ως ανταμοιβή την ακόλουθη:

$$r_i^7(t) = \left(\tilde{L}(t) + \sum_{j=1}^{N_B} B_j(t) - P_{disp}(t) \right)^2 + w_f \cdot \sum_{j=1}^{N_B} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2 + \psi_i(t) + \Lambda_i(t)$$

Όπου οι συναρτήσεις $\psi_i(t)$ και $\Lambda_i(t)$ υπολογίζονται με χρήση των σχέσεων (4.55) και (4.56) αντίστοιχα, ενώ το βάρος w_f αφήνεται ως παράμετρος προς βελτιστοποίηση.

Αρχικά, εξετάζεται η χρήση αυτής της μεθόδου χρησιμοποιώντας την αναπαράσταση της κατάστασης που χρησιμοποιεί την εκτίμηση του φορτίου, δηλαδή είναι:

$$s^7(t) = (SoE_1(t), \dots, SoE_{N_B}(t), \tilde{L}(t), t)$$

Με τον κάθε πράκτορα να λαμβάνει την παρατήρηση:

$$o_i^7(t) = (SoE_i(t), \tilde{L}(t), t)$$

Μέθοδος 8

Η τελευταία μέθοδος που εξετάζεται είναι η αντίστοιχη της μεθόδου 7 για την αναπαράσταση που χρησιμοποιεί τη συνολική ισχύ στο PCC, για την ενημέρωση των πρακτόρων. Έτσι, οι ζητούμενες συναρτήσεις που χρησιμοποιούνται είναι οι:

$$s^8(t) = (SoE_1(t), B_1(t-1), \dots, SoE_{N_B}(t), B_{N_B}(t-1), P_{PCC}(t-1), t)$$

$$o_i^8(t) = (SoE_i(t), B_i(t-1), P_{PCC}(t-1), t)$$

$$r_i^8(t) = r_i^7(t)$$

5.4 Επιλογή Παραμέτρων

Στην ενότητα αυτή παρουσιάζονται οι κύριες παράμετροι που χρησιμοποιήθηκαν στην εργασία και είναι κοινές για όλες τις μεθόδους. Ακόμη, παρουσιάζεται όπου κρίνεται χρήσιμο η κατεύθυνση για την κατάλληλη επιλογή της τιμής της παραμέτρου. Οι τιμές των υπόλοιπων παραμέτρων παρουσιάζονται στο [Παράρτημα Γ](#), εκτός αν εξαρτώνται από την περίπτωση, οπότε και γίνονται γνωστές στην επόμενη υποενότητα.

Δομή Νευρωνικών Δικτύων

Για τα δίκτυα actor και critic χρησιμοποιήθηκε η δομή που χρησιμοποιήθηκε και στην εργασία [7], χωρίς να εξεταστούν εναλλακτικές.

Συγκεκριμένα, ως δίκτυο actor χρησιμοποιείται ένα feedforward network, το οποίο δέχεται ως είσοδο τις παρατηρήσεις O_i του πράκτορα i και αποτελείται από 2 ενδιάμεσα επίπεδα με 64 units το κάθε ένα και από επίπεδο εξόδου με έναν νευρώνα. Για τα ενδιάμεσα units ως συνάρτηση ενεργοποίησης χρησιμοποιείται η *ReLU*, ενώ στο επίπεδο εξόδου χρησιμοποιείται η συνάρτηση *tanh*.

Αντίστοιχα, για το δίκτυο critic το feedforward network δέχεται ως είσοδο τη συνολική κατάσταση και τις δράσεις όλων των πρακτόρων, περιέχει 3 κρυφά επίπεδα με το κάθε ένα να αποτελείται από 64 νευρώνες και ένα επίπεδο εξόδου με μοναδικό unit. Σε κάθε unit του δικτύου critic χρησιμοποιείται η συνάρτηση *ReLU* ως συνάρτηση ενεργοποίησης.

Στην εκπαίδευση χρησιμοποιήθηκε ο βελτιστοποιητής (optimizer) Adam για όλα τα δίκτυα.

Παράμετρος γ

Ο παράγοντας μείωσης (discount factor) γ αποτελεί σημαντική παράμετρο για την εκπαίδευση των πρακτόρων ενισχυτικής μάθησης. Η τιμή του πρέπει να ρυθμιστεί, ώστε να λαμβάνονται αρκετά υπόψιν οι μελλοντικοί περιορισμοί, για να ικανοποιούνται κατά την εκτέλεση. Έτσι, για μικρές τιμές του οι πολιτικές που μαθαίνονται μπορεί να παραβιάζουν

τους συνδυαστικούς περιορισμούς, ενώ για μεγάλες οι λύσεις απέχουν αρκετά από τις οριακές τιμές του περιορισμού, ακόμα και αν προκύπτουν να είναι σημαντικά χειρότερες. Αυτό οφείλεται στη λανθασμένη εκτίμηση που δίνεται και δεν ενθαρρύνεται η εξερεύνηση τέτοιων λύσεων. Στην εργασία αυτή, χρησιμοποιείται η τιμή $\gamma = 0.2$ για μεθόδους που δεν χρησιμοποιούν τροχιές, και η τιμή $\gamma = 0.5$ για μεθόδους βασισμένες σε τροχιές. Η υψηλότερη τιμή σε μεθόδους που χρησιμοποιούν τροχιές κρίθηκε χρήσιμη, καθώς αποτρέπει λύσεις που απομακρύνονται σημαντικά από τις τροχιές και τελικά τις αγνοούν.

Παράμετροι σχετικοί με τροχιές

Στον ορισμό της συνάρτησης ανταμοιβής που χρησιμοποιεί τροχιές ορίστηκαν δύο παράμετροι, οι οποίες πρέπει να ρυθμιστούν. Η πρώτη είναι ο συντελεστής του κόστους μπαταρίας a_P . Αρχικά, εξετάστηκε η τιμή $a_P = 1$, η οποία είναι θεωρητικά αρκετή για τον στόχο που τέθηκε στην [ενότητα 2.4](#). Όμως οι προσομοιώσεις έδειξαν ότι στην εφαρμογή της ενισχυτικής μάθησης στο πρόβλημα αυτό δεν είναι αρκετή, ώστε να ακολουθούνται ικανοποιητικά οι τροχιές και κατ' επέκταση να επιτυγχάνεται ο στόχος της δίκαιης χρήσης. Έπειτα, εξετάστηκε η τιμή $a_P = \frac{1}{\max_i \{B_i^{max}\}}$. Αυτή η τιμή δίνει για το κόστος μπαταρίας $a_P \cdot \frac{(B_i(t) - B_i^*(SoE_i(t), t))^2}{4 \cdot B_i^{max}}$ τιμές στο $[0, 1]$ με τουλάχιστον μίας μπαταρίας το κόστος να μπορεί να πάρει την τιμή 1. Αυτή η τάξη μεγέθους κρίθηκε η κατάλληλη και χρησιμοποιήθηκε για όλες τις μεθόδους που χρησιμοποιήθηκαν, καθώς επιτύχανε τη δίκαιη χρήση, ενώ επέτρεπε στην πολιτική να προσαρμόζεται στην εκτίμηση του φορτίου, δηλαδή δεν υπερπροσαρμόζοταν στην τροχιά.

Σχετικά με το στόχο d_{max} , που χρησιμοποιείται για τον υπολογισμό των βαρών, η τιμή του πρέπει να είναι σχετικά μικρή. Επίσης, εξαρτάται άμεσα από τη μέγιστη αβεβαιότητα $|\Delta P|_{max}$, η οποία για τις περιπτώσεις που εξετάστηκαν έπαιρνε τιμές στο διάστημα $[0.0315, 0.032]$. Η τιμή που χρησιμοποιήθηκε στην εργασία είναι η $d_{max} = 0.052$. Παρατηρήθηκε ότι τιμές κοντά στην ελάχιστη είχαν ως αποτέλεσμα να αγνοούνται οι τροχιές, ενώ μεγάλες τιμές είχαν ως αποτέλεσμα η πολιτική να υπερπροσαρμόζεται στην τροχιά.

5.5 Πειράματα

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα των προσομοιώσεων για τις μεθόδους που ορίστηκαν. Για καλύτερη σύγκριση τα αποτελέσματα είναι οργανωμένα ανά περίπτωση χρήσης. Σχετικά με τη μεθοδολογία, αφού βρέθηκαν οι κατάλληλες τιμές των παραμέτρων, εκτελέστηκαν 5 προσομοιώσεις για κάθε μέθοδο, με τα αποτελέσματα που αναφέρονται στη συνέχεια να προκύπτουν ως ο μέσος όρος. Τα αποτελέσματα αυτά έχουν προκύψει χρησιμοποιώντας ως πραγματικό φορτίο ένα ημιτονοειδές σήμα με πλάτος 0.285p.u. και περίοδο 30min.

Υπενθυμίζεται ότι οι μετρικές που θα αξιολογηθούν είναι οι ακόλουθες:

- $cost = \sum_{t=k}^{T-1} (L^{real}(t) + \sum_{i=1}^{N_B} B_i(t) - P_{disp}(t))^2$
- $fairness_metric_1 = \frac{1}{T-k} \cdot \frac{1}{\binom{N_B}{2}} \cdot \frac{1}{4} \cdot \sum_{t=k}^{T-1} \sum_{i=1}^{N_B} \sum_{j=1}^{N_B} \left(\frac{B_i(t)}{B_i^{max}} - \frac{B_j(t)}{B_j^{max}} \right)^2$
- $fairness_metric_2 = \frac{1}{\binom{N_B}{2}} \sum_{i=1}^{N_B} \sum_{j \neq i} \left| \frac{SoE_i(T) - SoE_j(T)}{SoE_i(k) - SoE_j(k)} \right|$

Η πρώτη μετρική αφορά το κόστος δικτύου και όλες οι μέθοδοι πρέπει να έχουν μικρή τιμή. Η δεύτερη μετρική αφορά το κριτήριο δικαιοσύνης 2.5.1, επομένως οι μέθοδοι 3, 4, 7 και

8 πρέπει να δίνουν χαμηλή τιμή, για να χαρακτηριστεί ότι επιτυγχάνουν τον στόχο τους. Αντίστοιχα, η τρίτη μετρική σχετίζεται με το κριτήριο δικαιοσύνης 2.5.2, οπότε οι μέθοδοι 5 και 6 πρέπει να δίνουν χαμηλότερες τιμές από τις υπόλοιπες μεθόδους.

Επειδή η αναφορά των μεθόδων μέσω της αρίθμησης δεν είναι εύχρηστη και δεν δίνει άμεση διαίσθηση για τα χαρακτηριστικά της μεθόδου θα χρησιμοποιηθεί μία εναλλακτική περιγραφή τους. Δηλαδή κάθε μέθοδος θα δηλώνεται βάσει της αναπαράστασης και της ανταμοιβής της. Στον ακόλουθο πίνακα αναγράφονται οι αντιστοιχίσεις μεθόδων και περιγραφών.

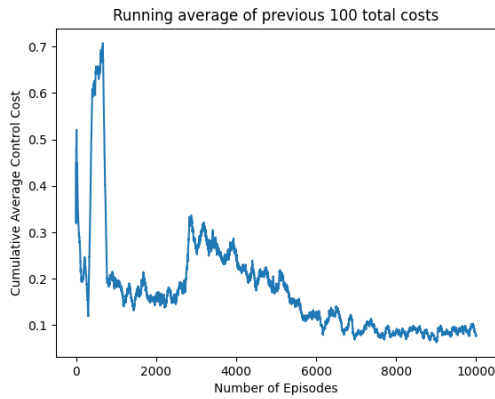
Μέθοδος	Κατάσταση	Ανταμοιβή
Μέθοδος 1	$(SoE_i(t), \tilde{L}(t), t)$	Βασική
Μέθοδος 2	$(SoE_i(t), P(t-1), B_i(t-1), t)$	Βασική
Μέθοδος 3	$(SoE_i(t), \tilde{L}(t), t)$	Τροχιές (2.5.1)
Μέθοδος 4	$(SoE_i(t), P(t-1), B_i(t-1), t)$	Τροχιές (2.5.1)
Μέθοδος 5	$(SoE_i(t), \tilde{L}(t), t)$	Τροχιές (2.5.2)
Μέθοδος 6	$(SoE_i(t), P(t-1), B_i(t-1), t)$	Τροχιές (2.5.2)
Μέθοδος 7	$(SoE_i(t), \tilde{L}(t), t)$	MSE
Μέθοδος 8	$(SoE_i(t), P(t-1), B_i(t-1), t)$	MSE

Στον παραπάνω πίνακα η κατάσταση δηλώνεται βάσει της παρατήρησης που λαμβάνει η κάθε μπαταρία για κάποια χρονική στιγμή t . Από την άλλη, η αναφορά ως ανταμοιβή δηλώνει ότι δεν χρησιμοποιείται κάποιο κριτήριο δικαιοσύνης, η λέξη τροχιές περιγράφει ότι η μέθοδος χρησιμοποιεί συνάρτηση ανταμοιβής βασισμένη σε τροχιές, ενώ στην παρένθεση δηλώνεται το κριτήριο δικαιοσύνης που χρησιμοποιεί η μέθοδος σχετικά με τη δίκαιη χρήση. Τέλος, με MSE δηλώνεται η συνάρτηση ανταμοιβής που ορίστηκε στην [ενότητα 4.5](#), η οποία χρησιμοποιεί το μέσο τετραγωνικό σφάλμα για να ενθαρρύνει τη δίκαιη χρήση.

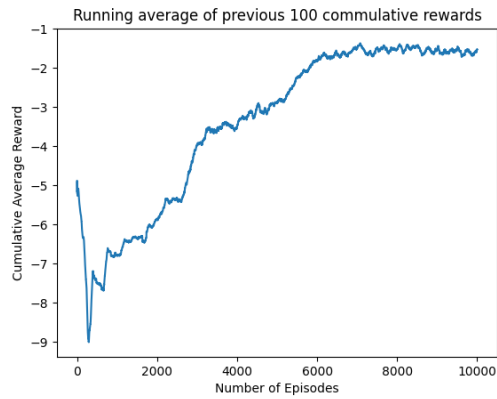
5.5.1 Περίπτωση 1

Αρχικά, εκτελούνται προσομοιώσεις για την πρώτη περίπτωση που παρουσιάστηκε, δηλαδή για την περίπτωση που το πλάνο διανομής προκύπτει από επίλυση ενός προβλήματος βελτιστοποίησης. Για την περίπτωση αυτή, στην εκπαίδευση εκτελέστηκαν 10000 επεισόδια για κάθε μέθοδο με το μοντέλο που προκύπτει να είναι αυτό με τη μεγαλύτερη μέση ανταμοιβή για τα τελευταία 50 επεισόδια. Ο χρόνος για την εκπαίδευση είναι περίπου 2 ώρες και 30 λεπτά για κάθε παράδειγμα. Ενδεικτικά, παρατίθεται η γραφική παράσταση για το κόστος και τη μέση ανταμοιβή που δίνουν τα μοντέλα για ένα παράδειγμα εκπαίδευσης στα σχήματα [5.3](#) και [5.4](#). Αντίστοιχα, ο χρόνος για να ληφθούν όλες οι αποφάσεις για όλες τις μπαταρίες και όλες τις χρονικές στιγμές, λάμβανε τιμές στο διάστημα $[0.8-1.2]$ sec, το οποίο οδηγεί σε χρόνο λήψης απόφασης πολύ μικρότερο του διαστήματος των 10sec, οπότε υπάρχει αρκετός χρόνος για την ενημέρωση των πρακτόρων από το PCC.

Πριν τα αποτελέσματα εξετάζονται τα κόστη που προκύπτουν με εφαρμογή μεθόδων που δεν χρησιμοποιούν ενισχυτική μάθηση. Η πρώτη μέθοδος είναι αυτή που προκύπτει λύση



Σχήμα 5.3: Μέσο Συνολικό Κόστος κατά την εκπαίδευση (1η περίπτωση)



Σχήμα 5.4: Μέση Συνολική Ανταμοιβή κατά την εκπαίδευση (1η περίπτωση)

με επίλυση του προβλήματος βελτιστοποίησης:

$$\begin{aligned}
 & \min_{B_1(t), B_2(t)} \sum_{t=0}^{T-1} [P_{PCC,d}(t) - P_{disp}(t)]^2 \\
 & \text{subject to } P_{PCC}(t) = \sum_{i=1}^2 B_i(t) + L_{real}(t) \\
 & \quad SoE_i(t+1) = SoE_i(t) + B_i(t)\Delta t \\
 & \quad B_i^{min} \leq B_i(t) \leq B_i^{max} \\
 & \quad SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \\
 & \quad SoE_1^{tr} + SoE_2^{tr} \leq SoE_1(T) + SoE_2(T) \leq SoE_A^{max,T} \\
 & \quad \text{for } i \in \{1, 2\}, t \in \{0, \dots, T-1\}
 \end{aligned}$$

Η μέθοδος αυτή θεωρεί εκ των προτέρων γνώση του φορτίου των μη ελεγχόμενων στοιχείων για όλο το διάστημα ελέγχου, επομένως δεν μπορεί να εφαρμοστεί στην πράξη. Έχει, όμως, θεωρητική αξία, καθώς δίνει ένα κάτω φράγμα για το κόστος των λύσεων που ικανοποιούν τους περιορισμούς.

Η δεύτερη μέθοδος που εξετάζεται είναι η κάθε μπαταρία να ακολουθεί την τροχιά της, χωρίς να λαμβάνει υπόψιν τις πληροφορίες για το σήμα σε πραγματικό χρόνο. Ιδανικά, ο στόχος του προβλήματος ενισχυτικής μάθησης είναι να δίνει καλύτερο αποτέλεσμα από αυτή τη μέθοδο, διαφορετικά θα ήταν ανούσια η εφαρμογή του.

Τελευταία μέθοδος που εξετάζεται είναι η εφαρμογή της μεθόδου της ενότητας 2.4. Για λόγους δίκαιης σύγκρισης χρησιμοποιήθηκαν οι τιμές των d_{max}, a_P που δώθηκαν και ως παράμετροι του προβλήματος ενισχυτικής μάθησης. Το μειονέκτημα αυτής της μεθόδου είναι ότι δεν λαμβάνει υπόψιν τον τελικό περιορισμό παρά μόνο την πληροφορία που δίνουν για αυτόν οι τροχιές, επομένως είναι πιθανό σε πολλές περιπτώσεις να μην ικανοποιείται ο περιορισμός.

Τα αποτελέσματα συγκεντρώνονται στον ακόλουθο πίνακα:

Κάτω Φράγμα	Τροχιές	Μέθοδος 2.4
0.000	0.008	0.005

Πίνακας 5.1: Αποτελέσματα κόστους μεθόδων αναφοράς για την περίπτωση 1

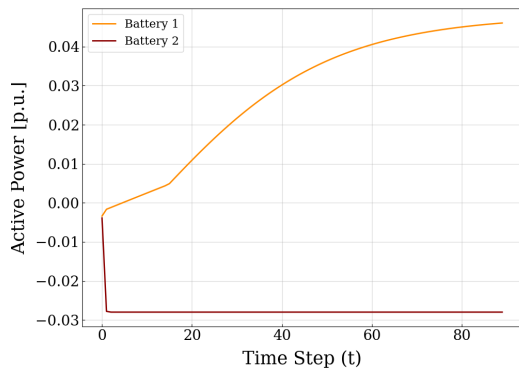
Μέθοδος 1

Πρώτα, εξετάστηκε η μέθοδος 1, στην οποία έπρεπε να ρυθμιστούν οι παράμετροι σχετικά με τις συναρτήσεις ποινής. Συγκεκριμένα, οι τιμές που επιλέχθηκαν είναι οι εξής:

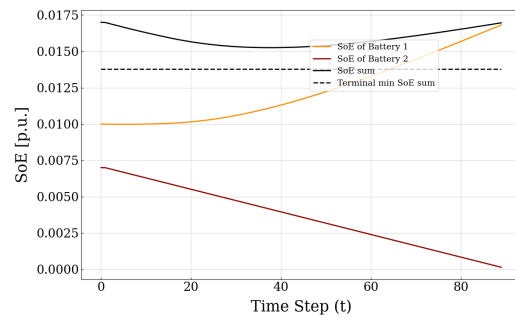
$$\psi_i(t) = \begin{cases} 0.5 & , SoE_i(t) < SoE_i^{min} \text{ or } SoE_i(t) > SoE_i^{max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 0.5 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) > SoE_A^{max,T} \\ 0.5 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) < SoE_1^{tr}(T) + SoE_2^{tr}(T) \\ 0.1 & , \text{if } t < T - 1 \text{ and } SoE_A^{max,T} - SoE_1(t+1) - SoE_2(t+1) < \sum_{k=1}^2 B_k^{min} \cdot \Delta T(t) \\ 0.1 & , \text{if } t < T - 1 \text{ and } SoE_1^{tr}(T) + SoE_2^{tr}(T) - SoE_1(t+1) - SoE_2(t+1) > \sum_{k=1}^2 B_k^{max} \cdot \Delta T(t) \\ 0 & , \text{else} \end{cases}$$

Ενδεικτικά, παρατίθεται η γραφική παράσταση για τις δράσεις των μπαταριών και της Κατάστασης της ενέργειας των μπαταριών από μία εκτέλεση:



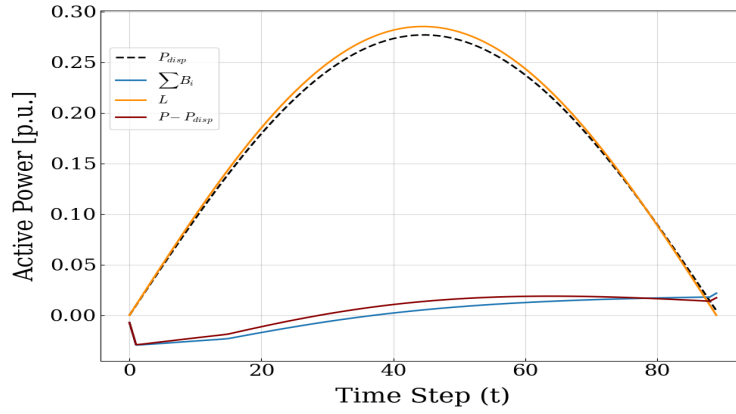
Σχήμα 5.5: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 1)



Σχήμα 5.6: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 1)

Αντίστοιχα, για την συνολική ενεργό ισχύ στο PCC και το αντίστοιχο πλάνο διανομής προκύπτει η γραφική παράσταση του σχήματος 5.7

Όπως προκύπτει από τα αποτελέσματα στη συγκεκριμένη περίπτωση, η μία μπαταρία φορτίζει, για να ικανοποιηθεί ο τελικός περιορισμός, ενώ η άλλη εκφορτίζει για να ακολουθεί το πλάνο διανομής, οπότε η χρήση δεν μπορεί να χαρακτηριστεί με κανένα από τα κριτήρια ως δίκαιη.



Σχήμα 5.7: Φορτίο και απόκλιση από το πλάνο διανομής στο PCC (Περίπτωση 1)

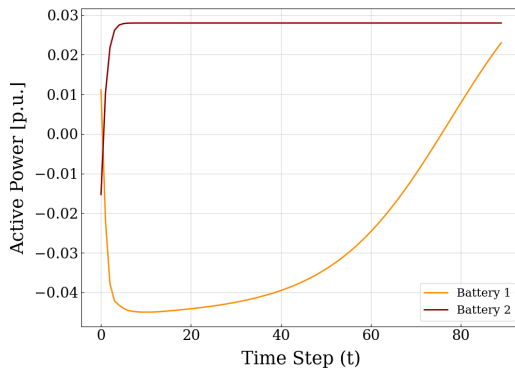
Μέθοδος 2

Και στην περίπτωση αυτή χρησιμοποιούνται οι συναρτήσεις ποινής:

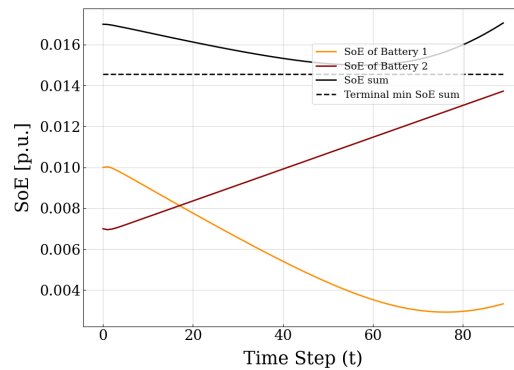
$$\psi_i(t) = \begin{cases} 0.5 & , \text{SoE}_i(t) < \text{SoE}_i^{\min} \text{ or } \text{SoE}_i(t) > \text{SoE}_i^{\max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 0.5 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) > \text{SoE}_A^{\max, T} \\ 0.5 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) < \text{SoE}_1^{\text{tr}}(T) + \text{SoE}_2^{\text{tr}}(T) \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_A^{\max, T} - \text{SoE}_1(t + 1) - \text{SoE}_2(t + 1) < \sum_{k=1}^2 B_k^{\min} \cdot \Delta T(t) \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_1^{\text{tr}}(T) + \text{SoE}_2^{\text{tr}}(T) - \text{SoE}_1(t + 1) - \text{SoE}_2(t + 1) > \sum_{k=1}^2 B_k^{\max} \cdot \Delta T(t) \\ 0 & , \text{else} \end{cases}$$

Οι αντίστοιχες γραφικές παραστάσεις για την ενεργό ισχύ που αποφασίζει η κάθε μπαταρία και την κατάσταση της ενέργειας της είναι οι ακόλουθες:



Σχήμα 5.8: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 2)



Σχήμα 5.9: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 2)

Ομοίως, και αυτή η μέθοδος δεν παρέχει δικαιοσύνη μεταξύ των μπαταριών.

Μέθοδος 3

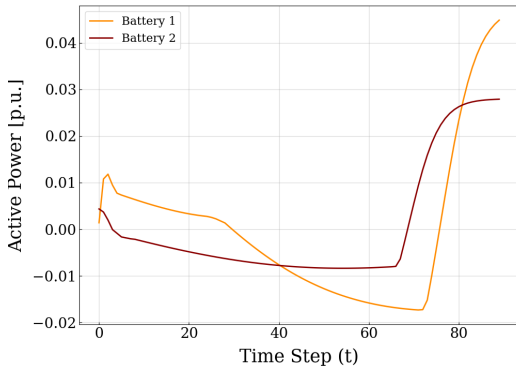
Η τρίτη μέθοδος είναι η πρώτη που εξετάζεται και επιχειρεί να δώσει λύση στο πρόβλημα της δίκαιης χρήσης. Επειδή οι τροχίες λειτουργούν ως proxy του τελικού περιορισμού η

συνάρτηση ποινής για τον τελικό περιορισμό δίνει ποινή μόνο για τη χρονική στιγμή που έχει οριστεί. Έτσι, για τη μέθοδο αυτή χρησιμοποιήθηκαν οι συναρτήσεις ποινής:

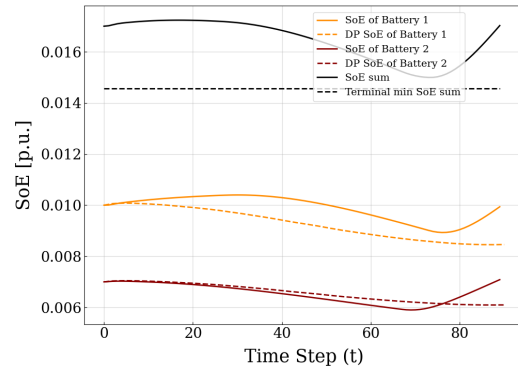
$$\psi_i(t) = \begin{cases} 0.5 & , SoE_i(t) < SoE_i^{min} \text{ or } SoE_i(t) > SoE_i^{max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 0.5 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) > SoE_A^{max,T} \\ 0.5 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) < SoE_1^{tr}(T) + SoE_2^{tr}(T) \\ 0 & , \text{else} \end{cases}$$

Σχετικά με τις δράσεις των μπαταριών, αλλά και τη κατάσταση της ενέργειάς τους κατά την περίοδο της εκτέλεσης προκύπτουν οι ακόλουθες γραφικές παραστάσεις, όπου με διακεκομμένες γραμμές αναπαρίστανται οι τροχιές:



Σχήμα 5.10: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 3)



Σχήμα 5.11: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 3)

Παρατηρείται ότι η ενεργός ισχύς των μπαταριών έχει το ίδιο πρόσημο για το μεγαλύτερο χρονικό διάστημα και οι γραφικές παραστάσεις των SoE έχουν παρόμοιες μορφές, το οποίο είναι επιθυμητό βάσει του κριτηρίου 2.5.1.

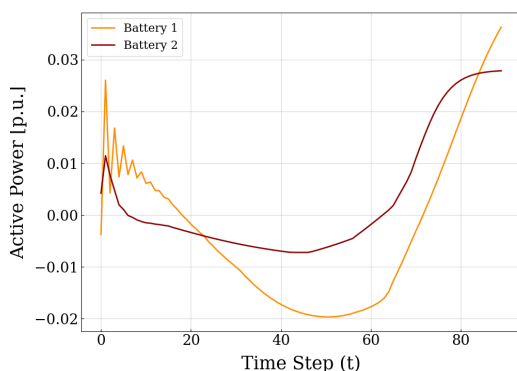
Μέθοδος 4

Η μέθοδος 4 διαφέρει με την προηγούμενη μόνο στην αναπαράσταση της κατάστασης που χρησιμοποιεί, γι' αυτό και χρησιμοποιείται η ίδια συνάρτηση ποινής. Από εκτέλεση αυτής της μεθόδου προκύπτουν οι γραφικές παραστάσεις των σχημάτων 5.12 και 5.13.

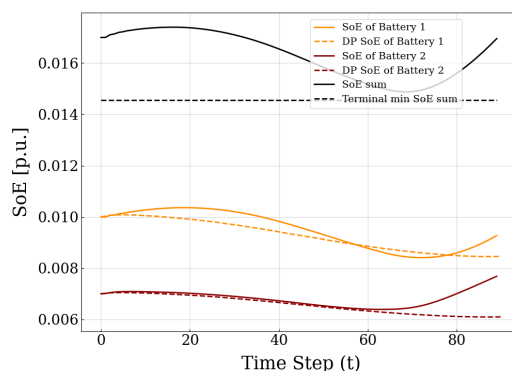
Όπως παρατηρείται από τις γραφικές παραστάσεις και αυτή η μέθοδος βελτιώνει σημαντικά τα αποτελέσματα όσον αφορά τη δικαιοσύνη μεταξύ των μπαταριών.

Μέθοδος 5

Η συνάρτηση ανταμοιβής που χρησιμοποιεί η μέθοδος αυτή είναι ίδια με τις προηγούμενες δύο, με την διαφορά να έγκειται στις τροχιές που χρησιμοποιούνται. Για αυτό τον λόγο χρησιμοποιήθηκε η ίδια συνάρτηση ποινής. Για τη μέθοδο αυτή ο στόχος είναι το κριτήριο δικαιοσύνης που παρουσιάστηκε στην ενότητα 2.5.2. Έτσι, ο στόχος είναι η σύγκλιση για τις τιμές του SoE των δύο μπαταριών. Σε μία από τις εκτελέσεις προέκυψαν οι γραφικές παραστάσεις, αναφορικά με την ενεργό ισχύ και την κατάσταση της ενέργειας των μπαταριών, που απεικονίζονται στα σχήματα 5.14 και 5.15. Όπως προκύπτει από τις γραφικές παραστάσεις το τελικό αποτέλεσμα είναι σημαντικά καλύτερο από τις μεθόδους 1 και 2, αναφορικά με το

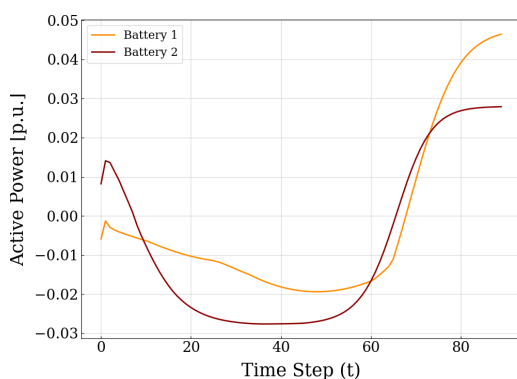


Σχήμα 5.12: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 4)

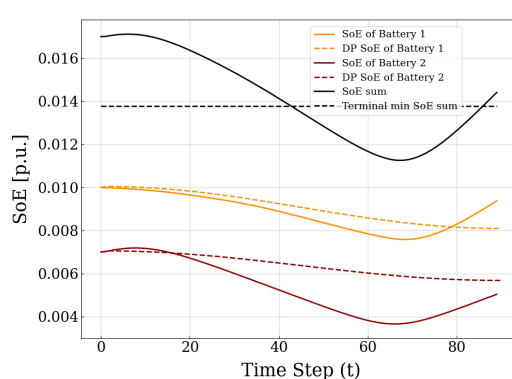


Σχήμα 5.13: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 4)

στόχο της δικαιοσύνης.



Σχήμα 5.14: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 5)



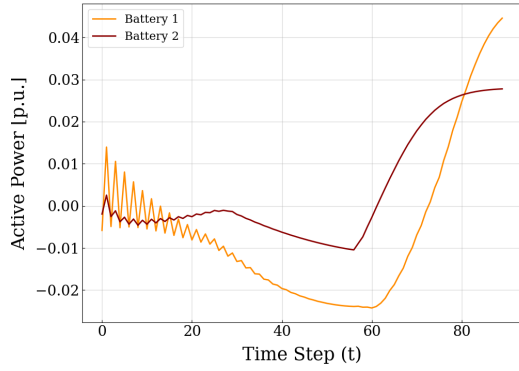
Σχήμα 5.15: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 5)

Μέθοδος 6

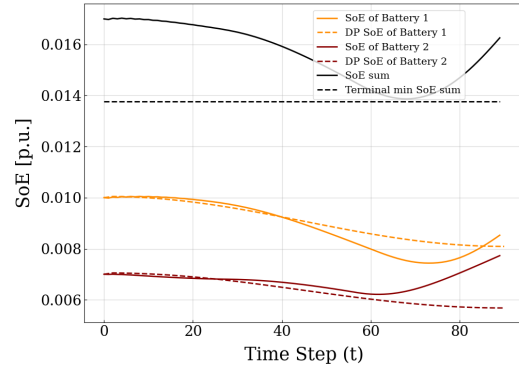
Η έκτη μέθοδος είναι αντίστοιχη της 5 με τη διαφορά ότι η κατάσταση που χρησιμοποιείται είναι αυτή με τη συνολική ισχύ στο PCC, έχοντας την ίδια συνάρτηση ανταμοιβής, οπότε χρησιμοποιείται και η ίδια συνάρτηση ποινής. Με χρήση αυτής τη μεθόδου προκύπτουν από μία εκτέλεση οι γραφικές παραστάσεις 5.16 και 5.17 για την ενεργό ισχύ και την κατάσταση της ενέργειας κάθε μπαταρίας αντίστοιχα. Από τη γραφική παράσταση φαίνεται ότι ο στόχος να συγκλίνουν οι τιμές του SoE ικανοποιείται.

Μέθοδος 7

Η έβδομη μέθοδος επιχειρούσε να προσφέρει δικαιοσύνη σύμφωνα με το κριτήριο 2.5.1, εισάγοντας άμεσα την απόκλιση από τη δίκαιη χρήση στη συνάρτηση ανταμοιβής. Σημαντικός παράμετρος είναι το βάρος w_f που δίνεται στο σφάλμα από τη δίκαιη χρήση. Στην περίπτωση 1 η τιμή που παρατηρήθηκε πειραματικά ότι εξισορροπεί τους δύο στόχους και δίνει καλά αποτελέσματα τόσο ως προς τον στόχο όσο και ως προς τη δικαιοσύνη είναι η $w_f = \frac{1}{400}$.



Σχήμα 5.16: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 6)



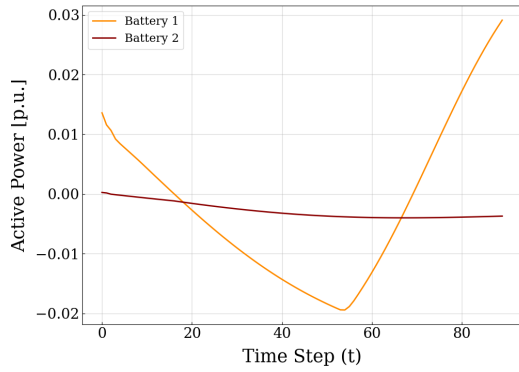
Σχήμα 5.17: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 6)

Αντίστοιχα, οι συναρτήσεις ποινής που χρησιμοποιήθηκαν είναι οι:

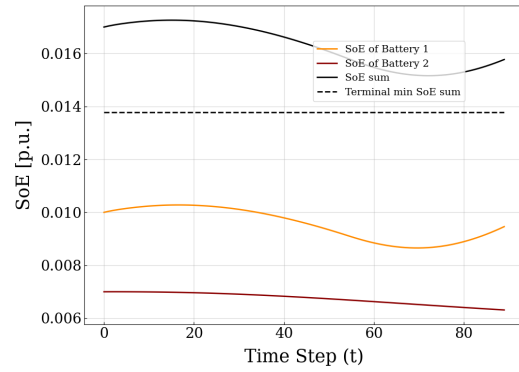
$$\psi_i(t) = \begin{cases} 0.5 & , \text{SoE}_i(t) < \text{SoE}_i^{\min} \text{ or } \text{SoE}_i(t) > \text{SoE}_i^{\max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 0.1 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) > \text{SoE}_A^{\max, T} \\ 0.1 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) < \text{SoE}_1^{\text{tr}}(T) + \text{SoE}_2^{\text{tr}}(T) \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_A^{\max, T} - \text{SoE}_1(t + 1) - \text{SoE}_2(t + 1) < \sum_{k=1}^2 B_k^{\min} \cdot \Delta T(t) \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_1^{\text{tr}}(T) + \text{SoE}_2^{\text{tr}}(T) - \text{SoE}_1(t + 1) - \text{SoE}_2(t + 1) > \sum_{k=1}^2 B_k^{\max} \cdot \Delta T(t) \\ 0 & , \text{else} \end{cases}$$

Από την εκτέλεση της μεθόδου αυτής, προέκυψαν οι ακόλουθες γραφικές παραστάσεις:



Σχήμα 5.18: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 7)



Σχήμα 5.19: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 7)

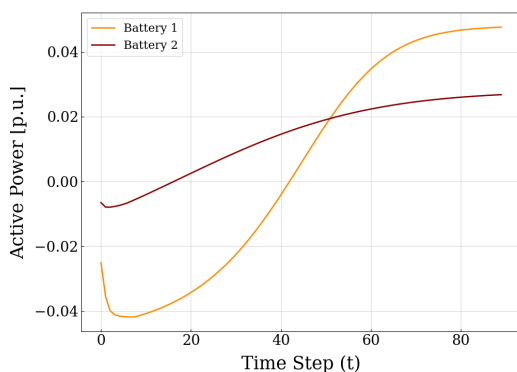
Για τη συγκεκριμένη περίπτωση, η μέθοδος αυτή δίνει βελτίωση σε σχέση με την αρχική, βάσει της μετρικής, όμως παρέχονται αποτελέσματα όπως το παραπάνω στο οποίο πρακτικά η μία μπαταρία δεν χρησιμοποιείται, το οποίο δεν είναι το δικαιότερο δυνατό.

Μέθοδος 8

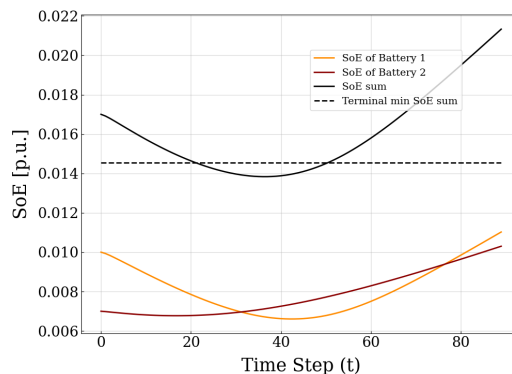
Η τελευταία μέθοδος που εξετάζεται είναι η εφαρμογή της ίδιας συνάρτησης ανταμοιβής που χρησιμοποιεί και η μέθοδος 7 με χρήση της αναπαράστασης της κατάστασης που χρη-

σιμοποιεί την ισχύ στο PCC. Οι συναρτήσεις ποινής, καθώς και το βάρος w_f έχουν τις ίδιες τιμές με αυτές που δόθηκαν στη μέθοδο 7.

Ενδεικτικά, παρατίθεται μία εκτέλεση αυτής της μεθόδου, η οποία δίνει αρκετά καλύτερα αποτελέσματα από τα αρχικά όσον αφορά τη δίκαιη χρήση, ενώ αποφεύγεται το πρόβλημα που φάνηκε στην προηγούμενη μέθοδο, που η μία μπαταρία δεν χρησιμοποιούταν. Οι γραφικές παραστάσεις για την ενεργό ισχύ και την κατάσταση της ενέργειας από αυτή την εκτέλεση φαίνονται στις ακόλουθες γραφικές παραστάσεις:



Σχήμα 5.20: Ενεργός ισχύς μπαταριών (Περίπτωση 1, Μέθοδος 8)



Σχήμα 5.21: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 1, Μέθοδος 8)

Σύγκριση αποτελεσμάτων

Τα αποτελέσματα που παρουσιάστηκαν έως τώρα προέκυπταν από μία μόνο εκτέλεση του αλγορίθμου. Όμως δεδομένου της τυχαιότητας που χρησιμοποιείται κατά την εκπαίδευση, μόνο μία εκτέλεση δεν είναι αρκετή, για να υποδείξει την ακρίβεια του μοντέλου. Γι' αυτό τον λόγο εκπαιδεύτηκαν 5 μοντέλα με κάθε μέθοδο που περιγράφηκε και στο σημείο αυτό θα παρουσιαστεί ο μέσος όρος των αποτελεσμάτων για κάθε μετρική που έχει οριστεί.

Αρχικά, παρουσιάζεται η μετρική του κόστους δικτύου στον ακόλουθο πίνακα :

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	0.047	0.049	0.114	0.213
$(SoE(t), P(t-1), B(t-1), t)$	0.046	0.044	0.040	0.207

Πίνακας 5.2: Αποτελέσματα μεθόδων MADDPG για το κόστος (Περίπτωση 1)

Τα συμπεράσματα που προκύπτουν είναι ότι η αναπαράσταση που χρησιμοποιεί τη συνολική ισχύ στο PCC είναι καλύτερη από την αναπαράσταση με το φορτίο των μη ελεγχόμενων στοιχείων L , καθώς για την ίδια συνάρτηση ανταμοιβής οδηγεί πάντα σε χαμηλότερο μέσο κόστος. Ακόμη, παρατηρείται ότι κάποιες μέθοδοι που χρησιμοποιούνται για τη δίκαιη χρήση καταφέρνουν να δώσουν αποτελέσματα για το κόστος δικτύου παρόμοια με τις μεθόδους 1, 2, δηλαδή δεν υπονομεύεται ο στόχος της δίκαιης χρήσης προς τον στόχο επίτευξης της βελτιστότητας. Τέλος, παρατηρείται ότι οι μέθοδοι αυτοί δίνουν αποτελέσματα αρκετά χειρότερα από το να ακολουθούσαν οι τροχιές, χωρίς να λαμβάνεται υπόψιν το πραγματικό φορτίο. Αυτό οφείλεται ότι για το σήμα που εξετάζεται οι τροχιές δίνουν κόστος κοντά στο βέλτιστο, οπότε σε τέτοιες περιπτώσεις είναι δύσκολη η εύρεση καλύτερης λύσης. Όμως,

υπάρχουν σήματα στα οποία οι μέθοδοι ενισχυτικής μάθησης δίνουν καλύτερα αποτελέσματα.

Στη συνέχεια, εξετάζεται η μετρική $fairness_metric_1$, η οποία αξιολογεί το κριτήριο δικαιοσύνης 2.5.1. Οι μέθοδοι που στοχεύουν στο να ικανοποιείται αυτό το κριτήριο είναι αυτές που χρησιμοποιούν τροχιές, υπολογισμένες σύμφωνα με αυτό το κριτήριο, και οι μέθοδοι που χρησιμοποιούν το μέσο τετραγωνικό σφάλμα (MSE) απευθείας στην ανταμοιβή. Αυτές οι μέθοδοι αντιστοιχούν στην τρίτη και πέμπτη στήλη του πίνακα. Ακολουθεί ο πίνακας με τη μέση τιμή του $fairness_metric_1$ έπειτα από 5 εκτελέσεις με κάθε μέθοδο:

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	0.848	0.102	0.251	0.254
$(SoE(t), P(t-1), B(t-1), t)$	0.844	0.082	0.088	0.124

Πίνακας 5.3: Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.1 (Περίπτωση 1)

Παρατηρείται ότι η μέθοδος που χρησιμοποιεί τροχιές βάσει του κριτηρίου 2.5.1 είναι αυτή που αποδίδει καλύτερα όσον αφορά τη δίκαιη χρήση. Η μέθοδος με το MSE στην ανταμοιβή παρέχει σημαντική βελτίωση σε σχέση με το να μην χρησιμοποιείται κανένα κριτήριο, όμως δίνει αποτελέσματα συγκρίσιμα με αυτά της μεθόδου με τροχιές, οι οποίες όμως έχουν σχεδιαστεί βάσει του κριτηρίου 2.5.2. Τέλος, παρατηρείται ότι και σε αυτή τη μετρική η αναπαράσταση $o_i(t) = (SoE_i(t), B_i(t-1), P(t-1), t)$ οδηγεί σε καλύτερα αποτελέσματα.

Τέλος, παρατίθενται αποτελέσματα για τη μετρική $fairness_metric_2$, η οποία αξιολογεί τις μεθόδους αναφορικά με τον στόχο να συγκλίνουν οι τιμές των SoE των μπαταριών. Οι μέθοδοι που έχουν σχεδιαστεί με αυτόν τον στόχο είναι αυτές που χρησιμοποιούν τροχιές, οι οποίες σχεδιάστηκαν βάσει του κριτηρίου δικαιοσύνης 2.5.2. Τα αποτελέσματα που προκύπτουν για αυτή τη μετρική παρουσιάζονται στον επόμενο πίνακα:

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	6.558	1.153	2.369	2.319
$(SoE(t), P(t-1), B(t-1), t)$	5.675	1.022	0.866	0.713

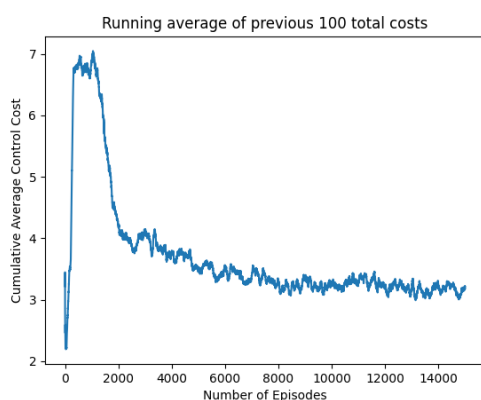
Πίνακας 5.4: Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.2 (Περίπτωση 1)

Όπως προκύπτει οι μέθοδοι αυτοί βελτιώνουν σημαντικά την τιμή της μετρικής σε σχέση με την περίπτωση που δεν χρησιμοποιείται κανένα κριτήριο δικαιοσύνης. Όμως, παρατηρείται ότι η τιμή της μετρικής σε μεθόδους που στοχεύουν στο κριτήριο δικαιοσύνης 2.5.1 είναι μικρότερη από αυτές που στοχεύουν άμεσα στη μετρική 2.5.2. Αυτό, όμως, δικαιολογείται, αφού όπως φαίνεται στις αντίστοιχες γραφικές παραστάσεις η μπαταρία 1, για την οποία ισχύει $B_1^{min} < B_2^{min}$ έχει αρχικά μεγαλύτερη τιμή SoE, δηλαδή $SoE_1(0) > SoE_2(0)$. Στην εκτέλεση προκύπτει για το σήμα που εξετάζεται για τις περισσότερες χρονικές στιγμές και οι δύο μπαταρίες να εκφορτίζουν για το μεγαλύτερο χρονικό διάστημα με τιμή κοντά στην ελάχιστη, με αποτέλεσμα να συγκλίνουν οι τιμές του SoE.

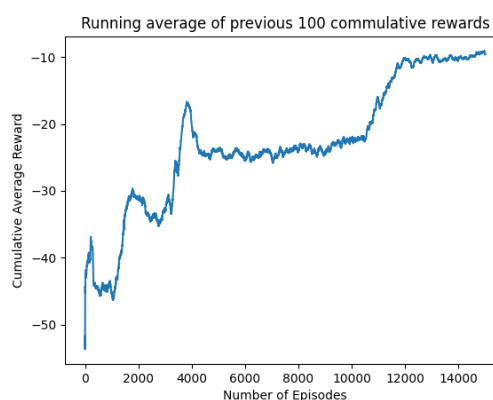
5.5.2 Περίπτωση 2

Η δεύτερη περίπτωση που εξετάζεται είναι αυτή που το πλάνο διανομής είναι ίσο με μηδέν για κάθε χρονική στιγμή. Η σημαντικότερη διαφορά σε σχέση με την προηγούμενη περίπτωση είναι ότι για τις περισσότερες χρονικές στιγμές στην περίπτωση αυτή είναι ανέφικτο το να ακολουθείται ακριβώς το πλάνο διανομής.

Από εκτελέσεις των προσομοιώσεων κρίθηκε ότι για την περίπτωση αυτή η εκτέλεση 15000 επεισοδίων κατά την εκπαίδευση οδηγεί σε καλύτερα αποτελέσματα, γι' αυτό και χρησιμοποιήθηκε αυτό το πλήθος. Ο χρόνος εκπαίδευσης σε αυτή την περίπτωση είναι περίπου 4 ώρες. Ενδεικτικά, παρατίθεται η γραφική παράσταση για το συνολικό κόστος και τη συνολική ανταμοιβή που παρατηρήθηκε κατά την εκπαίδευση ενός παραδείγματος για αυτή την περίπτωση:



Σχήμα 5.22: Μέσο Συνολικό Κόστος κατά την εκπαίδευση (2η περίπτωση)



Σχήμα 5.23: Μέση Συνολική Ανταμοιβή κατά την εκπαίδευση (2η περίπτωση)

Αναφορικά με τον χρόνο λήψης αποφάσεων, αυτός εξαρτάται από τη δομή του δικτύου του actor, οπότε είναι ο ίδιος και δεν εξαρτάται από την περίσταση, οπότε για όλες τις περιστάσεις υπάρχει επαρκής χρόνος για την ενημέρωση των πρακτόρων.

Αντίστοιχα, με την πρώτη περίπτωση εξετάστηκαν οι μέθοδοι αναφοράς που περιγράφθηκαν, για τις οποίες προέκυψαν τα εξής αποτελέσματα για το κόστος δικτύου:

Κάτω Φράγμα	Τροχιές	Μέθοδος 2.4
2.178	2.186	2.146*

Πίνακας 5.5: Αποτελέσματα κόστους μεθόδων αναφοράς για την περίπτωση 2

Σημειώνεται ότι με τη μέθοδο 2.4 δεν ικανοποιούνταν ο τελικός περιορισμός.

Μέθοδοι χωρίς κριτήριο δικαιοσύνης (1-2)

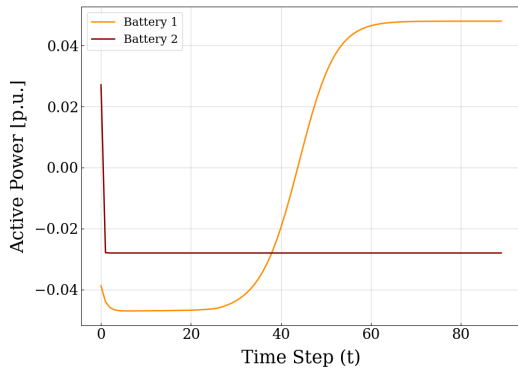
Πρώτα, εξετάστηκαν οι δύο μέθοδοι που χρησιμοποιούνται ως αναφορά, καθώς δεν στοχεύουν στη δίκαιη χρήση. Για την περίπτωση αυτή χρησιμοποιήθηκαν οι ακόλουθες συναρτήσεις ποινής:

$$\psi_i(t) = \begin{cases} 0.5 & , \text{SoE}_i(t) < \text{SoE}_i^{\min} \text{ or } \text{SoE}_i(t) > \text{SoE}_i^{\max} \\ 0 & , \text{else} \end{cases}$$

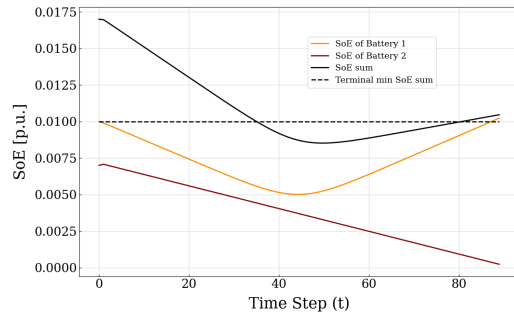
$$\Lambda_i(t) = \begin{cases} 3 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) > \text{SoE}_B^{\max,T} \\ 3 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) < \text{SoE}_B^{\min,T} \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_B^{\max,T} - \text{SoE}_1(t+1) - \text{SoE}_2(t+1) < \sum_{k=1}^2 B_k^{\min} \cdot \Delta T(t) \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_B^{\min,T} - \text{SoE}_1(t+1) - \text{SoE}_2(t+1) > \sum_{k=1}^2 B_k^{\max} \cdot \Delta T(t) \\ 0 & , \text{else} \end{cases}$$

Σε σχέση με την προηγούμενη περίπτωση η ποινή για τον συνδυαστικό περιορισμό παίρνει μεγαλύτερη τιμή. Αυτό κρίθηκε απαραίτητο, καθώς η προηγούμενη τιμή δεν ήταν αρκετή, ώστε να ικανοποιείται ο τελικός περιορισμός, ενώ αν χρησιμοποιούταν αυτή η τιμή στην προηγούμενη περίπτωση το κόστος δικτύου αυξανόταν σημαντικά.

Σχετικά με τις δράσεις των μπαταριών για τις μεθόδους αυτές και την κατάσταση της ενέργειας των μπαταριών, παρατίθενται οι ακόλουθες γραφικές παραστάσεις:



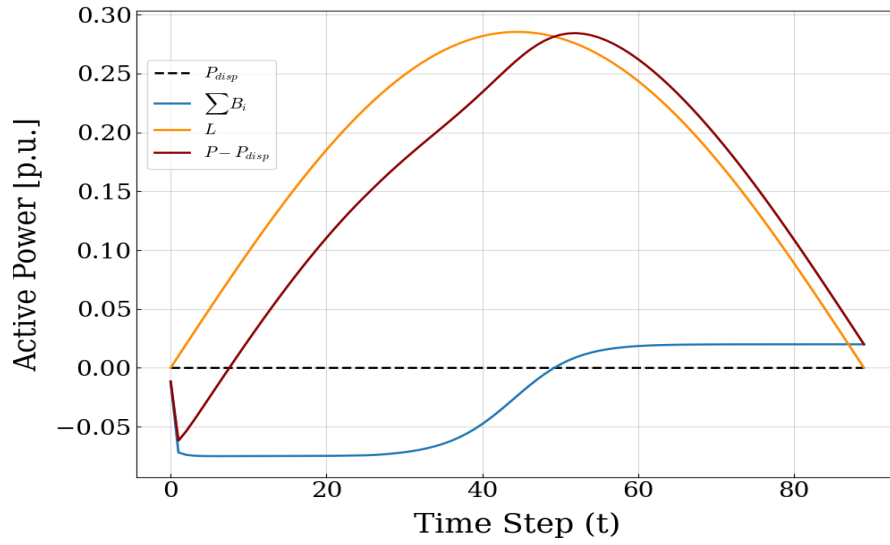
Σχήμα 5.24: Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 1)



Σχήμα 5.25: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 1)

Από τις γραφικές παραστάσεις ένα πρόβλημα που προέκυπτε σε πολλές εκτελέσεις αυτών των μεθόδων. Μετά από μία χρονική στιγμή η μία μπαταρία φόρτιζε για να ικανοποιηθεί ο τελικός περιορισμός, ενώ η άλλη εκφόρτιζε, καθώς εξακολουθούσε να στοχεύει στη μείωση του κόστους δικτύου, με αποτέλεσμα να αδειάζει.

Από την ίδια εκτέλεση, για την συνολική ισχύ που παρατηρούνταν στο PCC προέκυψε η γραφική παράσταση του σχήματος 5.26. Η μορφή είναι παρόμοια για όλες τις μεθόδους της περίπτωσης αυτής, οπότε δεν θα παρατεθεί για τις υπόλοιπες.



Σχήμα 5.26: Φορτίο και απόκλιση από το πλάνο διανομής στο PCC (Περίπτωση 2)

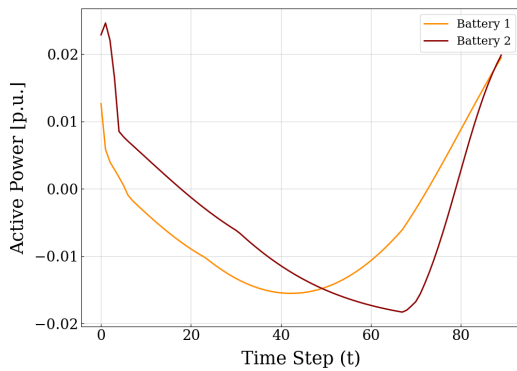
Μέθοδοι με τροχιές (3-6)

Σχετικά με τις μεθόδους που χρησιμοποιούν τροχιές οι συναρτήσεις ποινής που έδωσαν τα καλύτερα αποτελέσματα είναι οι εξής:

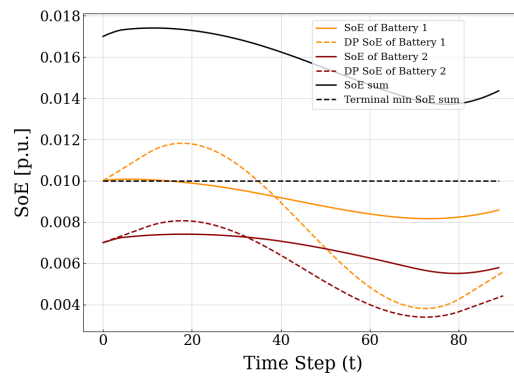
$$\psi_i(t) = \begin{cases} 0.5 & , SoE_i(t) < SoE_i^{min} \text{ or } SoE_i(t) > SoE_i^{max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 4 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) > SoE_B^{max,T} \\ 4 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) < SoE_B^{min,T} \\ 0 & , \text{else} \end{cases}$$

Οι μέθοδοι 3 και 4 στοχεύουν στο κριτήριο δικαιοσύνης 2.5.1. Από ένα παράδειγμα μοντέλου που εκπαιδεύτηκε με τη μέθοδο 3 προέκυψαν οι ακόλουθες γραφικές παραστάσεις:

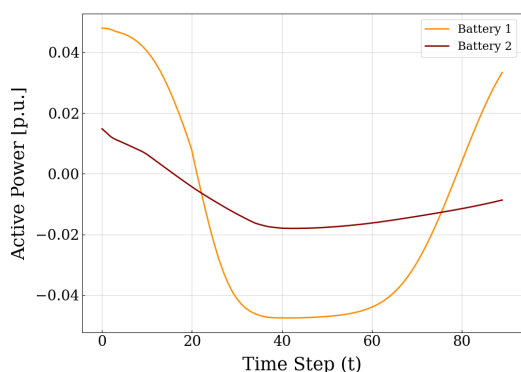


Σχήμα 5.27: Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 3)

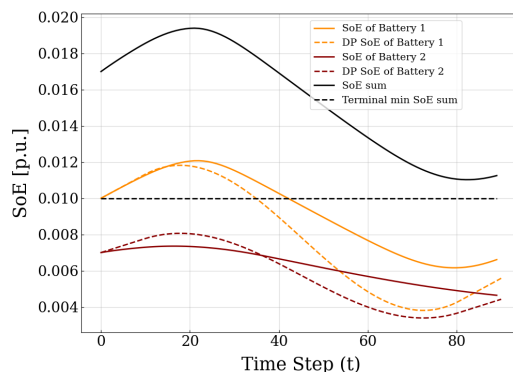


Σχήμα 5.28: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 3)

Από τις γραφικές παραστάσεις είναι εμφανές ότι η χρήση των μπαταριών είναι δικαιότερη από ότι σε αυτές της προηγούμενης μεθόδου. Σε ένα σημαντικό ποσοστό παρατηρούνταν το πρόβλημα να φορτίζει μόνο η μία στο τέλος, ώστε να ικανοποιείται ο τελικός περιορισμός. Αυτό φαίνεται στο ακόλουθο παράδειγμα που προέκυψε μέσω της μεθόδου 4:



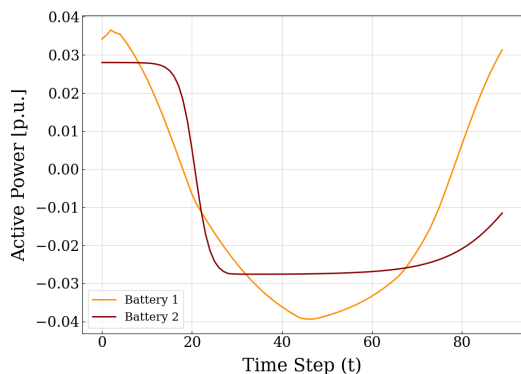
Σχήμα 5.29: Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 4)



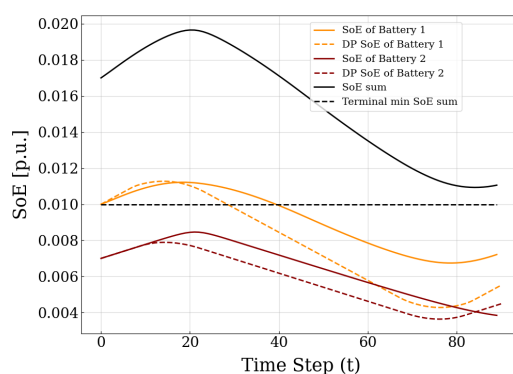
Σχήμα 5.30: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 4)

Ακόμη, όμως, και σε τέτοιες περιπτώσεις μέσω αυτών των μεθόδων το χρονικό διάστημα που η χρήση των μπαταριών είναι άδικη είναι σημαντικά συντομότερο, με αποτέλεσμα αυτή η μέθοδος να εξακολουθεί να είναι δικαιότερη, όπως θα επιβεβαιωθεί και μέσω της μετρικής $fairness_metric_1$, που θα παρουσιαστεί στη συνέχεια.

Αντίστοιχα, οι μέθοδοι 5 και 6 στοχεύουν στο κριτήριο δικαιοσύνης 2.5.2. Εκπαιδεύοντας ένα μοντέλο μέσω της μεθόδου 5, προέκυψαν κατά την αξιολόγηση οι ακόλουθες γραφικές παραστάσεις:



Σχήμα 5.31: Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 5)



Σχήμα 5.32: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 5)

Σε αντίθεση με την αρχική μέθοδο που στο τέλος οι τιμές του SoE απέκλιναν, μέσω αυτής της μεθόδου φαίνεται ότι αποτρέπονται τέτοια προβλήματα, το οποίο είναι ουσιαστικά ο στόχος αυτού του κριτηρίου.

Μέθοδοι με MSE (7-8)

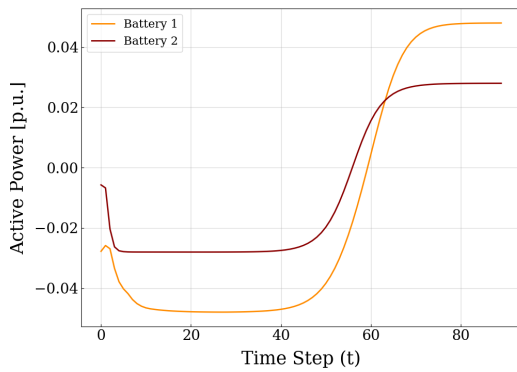
Τελευταίες μέθοδοι που εξετάστηκαν και για αυτή την περίπτωση είναι αυτές που εφαρμόζουν απευθείας στη συνάρτηση ανταμοιβής το κριτήριο δικαιοσύνης 2.5.1. Για αυτήν την

περίπτωση χρησιμοποιήθηκε η παράμετρος $\omega_f = \frac{1}{100}$. Η χρήση μεγαλύτερης τιμής για το βάρος που δίνεται στη δίκαιη χρήση είναι δικαιολογημένη, καθώς σε αυτή την περίπτωση οι τιμές για το κόστος δικτύου έχουν μεγαλύτερη τάξη μεγέθους, οπότε πρέπει να δοθεί μεγαλύτερο βάρος στο κριτήριο δικαιοσύνης, για να μην υπονομευθεί. Ως συναρτήσεις ποινής χρησιμοποιήθηκαν, επίσης, οι:

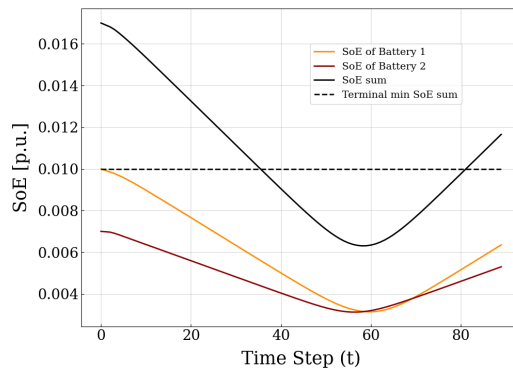
$$\psi_i(t) = \begin{cases} 0.5 & , SoE_i(t) < SoE_i^{min} \text{ or } SoE_i(t) > SoE_i^{max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 3 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) > SoE_B^{max,T} \\ 3 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) < SoE_B^{min,T} \\ 0.1 & , \text{if } t < T - 1 \text{ and } SoE_B^{max,T} - SoE_1(t+1) - SoE_2(t+1) < \sum_{k=1}^2 B_k^{min} \cdot \Delta T(t) \\ 0.1 & , \text{if } t < T - 1 \text{ and } SoE_B^{min,T} - SoE_1(t+1) - SoE_2(t+1) > \sum_{k=1}^2 B_k^{max} \cdot \Delta T(t) \\ 0 & , \text{else} \end{cases}$$

Ένα παράδειγμα από την αξιολόγηση μοντέλων που εκπαιδεύτηκαν με τη μέθοδο 8 φαίνεται στις ακόλουθες γραφικές παραστάσεις:



Σχήμα 5.33: Ενεργός ισχύς μπαταριών (Περίπτωση 2, Μέθοδος 8)



Σχήμα 5.34: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 2, Μέθοδος 8)

Από τις γραφικές παραστάσεις είναι εμφανές ότι η χρήση των μπαταριών είναι δίκαιη, σύμφωνα με το αντίστοιχο κριτήριο.

Σύγκριση αποτελεσμάτων

Αφού δόθηκαν τα πλεονεκτήματα των μεθόδων μέσω συγκεκριμένων παραδειγμάτων, στο μέρος αυτό θα αξιολογηθούν αυτές μέσω των αντικειμενικών μετρικών που ορίστηκαν. Τα αποτελέσματα προκύπτουν ως ο μέσος όρος από την αξιολόγηση της μεθόδου για 5 διαφορετικά παραδείγματα εκπαίδευσης, ώστε να πιστοποιηθεί ότι οι μέθοδοι δίνουν σταθερά καλά αποτελέσματα.

Αρχικά, εξετάζονται οι μέθοδοι στο πως αποδίδουν στον στόχο του δικτύου. Στον ακόλουθο πίνακα παρουσιάζεται το μέσο κόστος του δικτύου.

Παρατηρείται ότι και σε αυτή την περίπτωση οι μέθοδοι, που χρησιμοποιούν για την αναπαράσταση της κατάστασης τη συνολική ισχύ στο PCC, αποδίδουν καλύτερα σχετικά με τον στόχο δικτύου. Επίσης, οι μέθοδοι που χρησιμοποιούν τροχιές δίνουν βέλτιστα αποτελέσματα με τη μέση τιμή να είναι κοντά στο κόστος που δίνουν οι τροχιές, ενώ για άλλα

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	3.323	2.705	2.601	2.872
$(SoE(t), P(t-1), B(t-1), t)$	2.920	2.563	2.594	2.823

Πίνακας 5.6: Αποτελέσματα μεθόδων MADDPG για το κόστος (Περίπτωση 2)

φορτία δίνουν καλύτερα αποτελέσματα. Τέλος, οι μέθοδοι που χρησιμοποιούν απευθείας το κόστος για τη δίκαιη χρήση στην ανταμοιβή έχουν συγκρίσιμη απόδοση με τις μεθόδους με τροχιές σε αυτή την περίπτωση.

Ακολουθώς αξιολογούνται οι μέθοδοι ως προς τον στόχο δικαιοσύνης 2.5.1, σύμφωνα με τη μετρική *fairness_metric*₁. Οι μέθοδοι που έχουν σχεδιαστεί βάσει αυτού του κριτηρίου είναι αυτές που χρησιμοποιούν τροχιές, σχεδιασμένες βάσει του αντίστοιχου κριτηρίου και η μέθοδος που εισάγει απευθείας το μέσο τετραγωνικό σφάλμα στη συνάρτηση ανταμοιβής. Η μέση τιμή της μετρικής, όπως προέκυψε από 5 εκτελέσεις κάθε μεθόδου εμφανίζεται στον επόμενο πίνακα:

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	0.724	0.117	0.146	0.101
$(SoE(t), P(t-1), B(t-1), t)$	0.425	0.158	0.178	0.063

Πίνακας 5.7: Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.1 (Περίπτωση 2)

Βάσει των αποτελεσμάτων προκύπτει ότι η μέθοδος που λειτουργεί καλύτερα είναι αυτή που χρησιμοποιεί άμεσα το MSE. Η μέθοδος με τις τροχιές εξακολουθεί να δίνει αποτελέσματα, τα οποία είναι κοντά σχετικά στα βέλτιστα και σημαντικά καλύτερα από τις μεθόδους που δεν χρησιμοποιούν κανένα κριτήριο δικαιοσύνης. Σχετικά με τις μεθόδους που χρησιμοποιούν τροχιές, για την περίπτωση αυτή η αναπαράσταση $(SoE(t), \tilde{L}(t), t)$ δίνει πιο δίκαια αποτελέσματα, με τη διαφορά όμως να μην είναι ιδιαίτερα σημαντική.

Τέλος, εξετάζεται η μετρική *fairness_metric*₂, η οποία στοχεύει στο κριτήριο δικαιοσύνης 2.5.2. Οι τιμές για το μέσο όρο της μετρικής αυτής φαίνεται στον πίνακα που ακολουθεί:

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	3.807	1.472	1.321	0.599
$(SoE(t), P(t-1), B(t-1), t)$	3.130	1.245	1.593	0.588

Πίνακας 5.8: Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.2 (Περίπτωση 2)

Παρατηρείται ότι η μέθοδος με τροχιές που στοχεύουν σε αυτό το είδος δικαιοσύνης δίνουν σημαντικά καλύτερα αποτελέσματα από την αρχική μέθοδο. Όμως παρατηρείται ότι οι άλλες μέθοδοι δίνουν καλύτερα αποτελέσματα, αυτό όμως δικαιολογείται όπως και στην προηγούμενη περίπτωση και δεν αναμένεται να ισχύει για διαφορετικές περιπτώσεις.

5.5.3 Περίπτωση 3

Η τελευταία περίπτωση που εξετάζεται αφορά ένα δίκτυο, το οποίο αποτελείται από 4 μπαταρίες. Και σε αυτή την περίπτωση το πλάνο διανομής θεωρείται ίσο με το μηδέν για

κάθε χρονική στιγμή, όμως οι περισσότερες μπαταρίες αυξάνουν τις δυνατότητες του δικτύου, ώστε να ακολουθούνται τέτοιες περιπτώσεις.

Σχετικά με την εκπαίδευση, η εκτέλεση για 10000 επεισόδια κρίθηκε επαρκής, για να επιτευχθούν τα επιθυμητά αποτελέσματα. Ο χρόνος που χρειάστηκε είναι περίπου 5 ώρες, δηλαδή διπλάσιος σε σχέση με τον χρόνο που χρειάζεται για δίκτυο 2 μπαταριών για τον ίδιο αριθμό επεισοδίων. Η μεταβολή του κόστους και της συνολικής ανταμοιβής κατά την εκπαίδευση από ένα παράδειγμα αυτής της μεθόδου φαίνεται στις γραφικές παραστάσεις 5.35 και 5.36.

Όπως αναφέρθηκε, η αύξηση του πλήθους του μπαταριών αυξάνει και τις δυνατότητες αποθήκευσης και προσφοράς ενέργειας του δικτύου. Αυτό μπορεί να επιβεβαιωθεί και με την εξέταση των μεθόδων αναφοράς που εξετάστηκαν και στις προηγούμενες περιπτώσεις. Οι τιμές για το κόστος δικτύου για την περίπτωση αυτή παρουσιάζονται στον πίνακα που ακολουθεί:

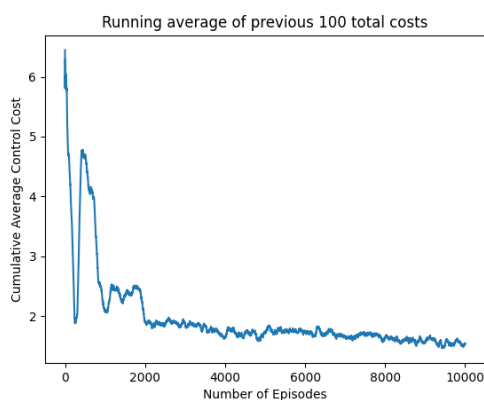
Κάτω Φράγμα	Τροχιές	Μέθοδος 2.4
0.951	0.967	0.944*

Πίνακας 5.9: Αποτελέσματα κόστους μεθόδων αναφοράς για την περίπτωση 3

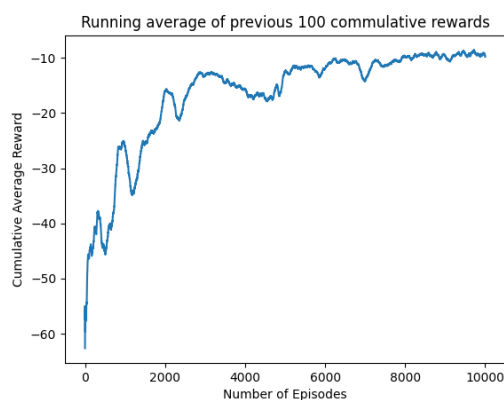
Η μέθοδος 2.4 δεν ικανοποιούσε τον τελικό περιορισμό, γι' αυτό και έχει τιμή μικρότερη από το κάτω φράγμα.

Μέθοδοι χωρίς κριτήριο δικαιοσύνης (1-2)

Όπως και για τις προηγούμενες περιπτώσεις εξετάζεται η εφαρμογή της αλγορίθμου MADDPG για τις 2 προτεινόμενες αναπαραστάσεις χωρίς να χρησιμοποιείται κάποιο κριτήριο δικαιοσύνης στην ανταμοιβή. Σχετικά με τις συναρτήσεις ποινής χρησιμοποιήθηκαν οι ακόλουθες:



Σχήμα 5.35: Μέσο Συνολικό Κόστος κατά την εκπαίδευση (3η περίπτωση)

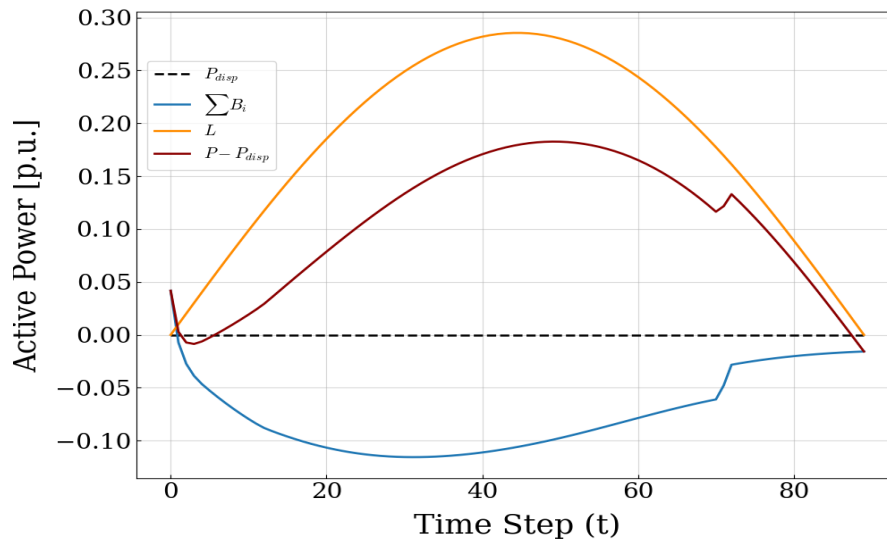


Σχήμα 5.36: Μέση Συνολική Ανταμοιβή κατά την εκπαίδευση (3η περίπτωση)

$$\psi_i(t) = \begin{cases} 0.5 & , \text{SoE}_i(t) < \text{SoE}_i^{\min} \text{ or } \text{SoE}_i(t) > \text{SoE}_i^{\max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 3 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) + \text{SoE}_3(T) + \text{SoE}_4(T) > \text{SoE}_C^{\max,T} \\ 3 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) + \text{SoE}_3(T) + \text{SoE}_4(T) < \text{SoE}_C^{\min,T} \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_C^{\max,T} - \sum_{k=1}^4 \text{SoE}_k(t+1) < \sum_{k=1}^4 B_k^{\min} \cdot \Delta T(t) \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_C^{\min,T} - \sum_{k=1}^4 \text{SoE}_k(t+1) > \sum_{k=1}^4 B_k^{\max} \cdot \Delta T(t) \\ 0 & , \text{else} \end{cases}$$

Σχετικά με τον στόχο του να ακολουθείται το πλάνο διανομής, στη γραφική παράσταση του σχήματος 5.37 φαίνεται η συνολική ισχύς στο PCC για το φορτίο που εξετάζεται. Όπως προκύπτει, η απόκλιση είναι μικρότερη σε σχέση με την περίπτωση 2, το οποίο είναι αναμενόμενο λόγω των περισσότερων μπαταριών.

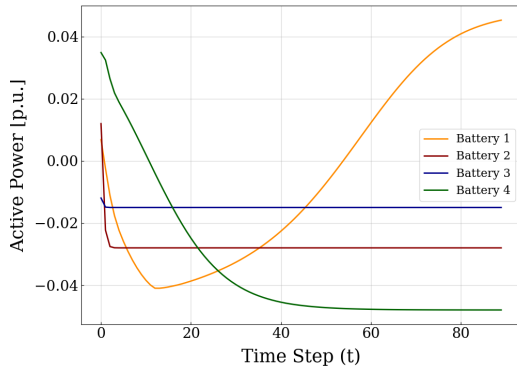


Σχήμα 5.37: Φορτίο και απόκλιση από το πλάνο διανομής στο PCC (Περίπτωση 3)

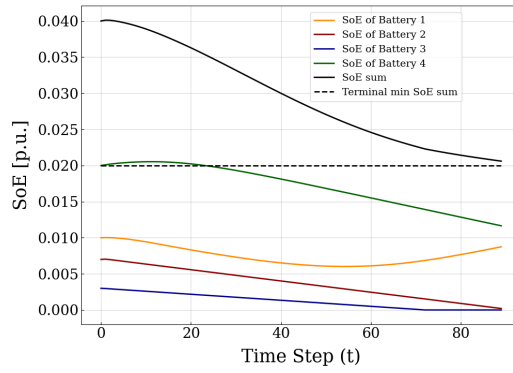
Στη συνέχεια, παρουσιάζονται στα σχήματα 5.38 και 5.39 οι γραφικές παραστάσεις για την ενεργό ισχύ κάθε μπαταρίας και την κατάσταση της ενέργειας της. Από αυτές προκύπτει ότι για το παράδειγμα που παρουσιάζεται η μία μπαταρία φορτίζει συνεχώς, μία άλλη εκφορτίζει για όλο το χρονικό διάστημα, ενώ οι άλλες δύο κάποιες χρονικές στιγμές φορτίζουν, ενώ κάποιες άλλες εκφορτίζουν. Επομένως, προκύπτει ότι η βασική συνάρτηση ανταμοιβής δεν είναι αρκετή για τη δίκαιη χρήση, οπότε εξετάζονται οι εναλλακτικές συναρτήσεις που προτάθηκαν, προκειμένου να εξεταστεί αν επιλύεται το πρόβλημα.

Μέθοδοι με τροχιές (3-6)

Η πρώτη προσπάθεια για να μαθαίνονται δίκαιες πολιτικές είναι με την εισαγωγή τροχιών για την κατάσταση της ενέργειας μέσω της συνάρτησης ανταμοιβής. Για την περίπτωση των 4 μπαταριών χρησιμοποιήθηκε η παρακάτω συνάρτηση ανταμοιβής:



Σχήμα 5.38: Ενεργός ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 2)

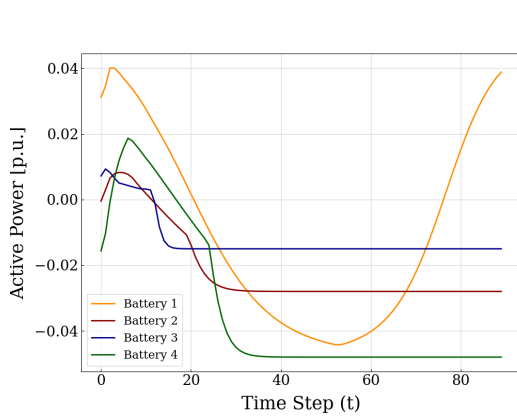


Σχήμα 5.39: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 2)

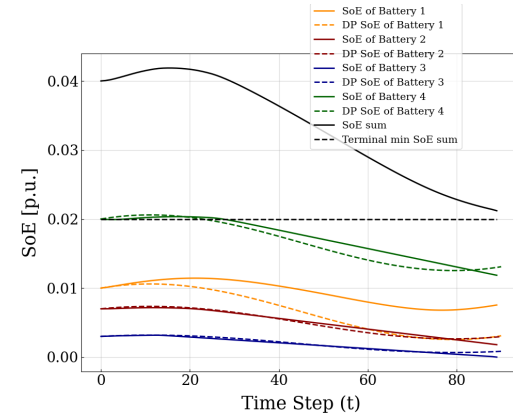
$$\psi_i(t) = \begin{cases} 0.5 & , SoE_i(t) < SoE_i^{min} \text{ or } SoE_i(t) > SoE_i^{max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 2 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) + SoE_3(T) + SoE_4(T) > SoE_C^{max,T} \\ 2 & , \text{if } t = T - 1 \text{ and } SoE_1(T) + SoE_2(T) + SoE_3(T) + SoE_4(T) < SoE_C^{min,T} \\ 0 & , \text{else} \end{cases}$$

Αρχικά, παρουσιάζεται η γραφική παράσταση για την ενεργό ισχύ και για την κατάσταση της ενέργειας από μία εκτέλεση για αξιολόγηση της μεθόδου 3.



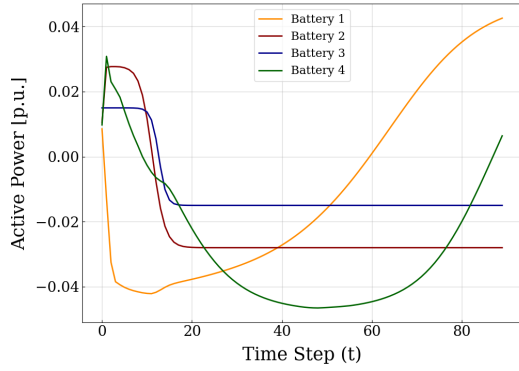
Σχήμα 5.40: Ενεργός ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 3)



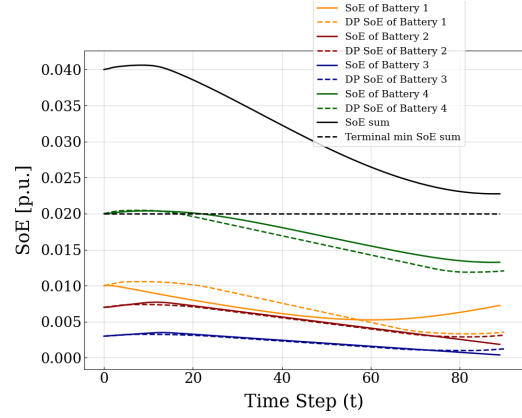
Σχήμα 5.41: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 3)

Η μέθοδος 3, όπως και η 4, στοχεύει στη δίκαιη χρήση βάσει του κριτηρίου 2.5.1. Βάσει των γραφικών παραστάσεων προκύπτει ότι, ουσιαστικά, για το μεγαλύτερο χρονικό διάστημα οι μπαταρίες χρησιμοποιούνται δίκαια, με μοναδική εξαίρεση τις τελευταίες χρονικές στιγμές, όπου η μπαταρία 1 αρχίζει να φορτίζει, για να ικανοποιείται ο τελικός περιορισμός, σε αντίθεση με τις υπόλοιπες.

Αντίστοιχες γραφικές παραστάσεις που προκύπτουν μέσω της μεθόδου 6 παρουσιάζονται ακολούθως.



Σχήμα 5.42: Ενεργός ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 6)



Σχήμα 5.43: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 6)

Η μέθοδος αυτή, όπως και η μέθοδος 5, στοχεύουν στο κριτήριο δικαιοσύνης 2.5.2. Από τις γραφικές παραστάσεις προκύπτει ότι οι τροχιές που παράγονται από πολιτική που έμαθε ο αντίστοιχος πράκτορας είναι αρκετά κοντά στην τροχιά που χρησιμοποιήθηκε κατά την εκπαίδευση, επομένως η μέθοδος συμβάλλει στη δίκαιη χρήση βάσει του κριτηρίου.

Μέθοδοι με MSE (7-8)

Τέλος, εξετάζονται οι δύο διαφορετικές αναπαραστάσεις σε συνδυασμό με τη συνάρτηση ανταμοιβής, η οποία χρησιμοποιεί ως ποινή το μέσο τετραγωνικό σφάλμα από τη δίκαιη χρήση. Στην περίπτωση των 4 μπαταριών η παράμετρος $\omega_f = \frac{1}{400}$ κρίθηκε ότι είναι η κατάλληλη τιμή για να εξισορροπούνται οι δύο στόχοι. Αντίστοιχα, οι συναρτήσεις ποινής που χρησιμοποιήθηκαν είναι οι:

$$\psi_i(t) = \begin{cases} 0.5 & , \text{SoE}_i(t) < \text{SoE}_i^{\min} \text{ or } \text{SoE}_i(t) > \text{SoE}_i^{\max} \\ 0 & , \text{else} \end{cases}$$

$$\Lambda_i(t) = \begin{cases} 3 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) + \text{SoE}_3(T) + \text{SoE}_4(T) > \text{SoE}_C^{\max, T} \\ 3 & , \text{if } t = T - 1 \text{ and } \text{SoE}_1(T) + \text{SoE}_2(T) + \text{SoE}_3(T) + \text{SoE}_4(T) < \text{SoE}_C^{\min, T} \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_C^{\max, T} - \sum_{k=1}^4 \text{SoE}_k(t+1) < \sum_{k=1}^4 B_k^{\min} \cdot \Delta T(t) \\ 0.1 & , \text{if } t < T - 1 \text{ and } \text{SoE}_C^{\min, T} - \sum_{k=1}^4 \text{SoE}_k(t+1) > \sum_{k=1}^4 B_k^{\max} \cdot \Delta T(t) \\ 0 & , \text{else} \end{cases}$$

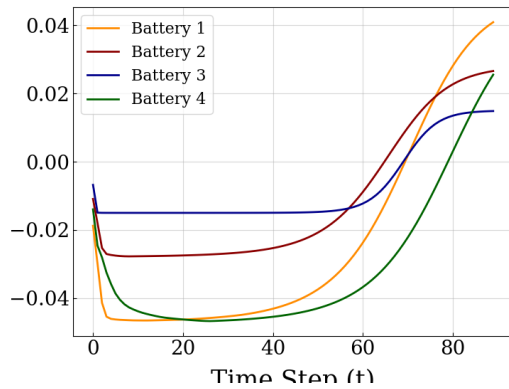
Οι γραφικές παραστάσεις για την ενεργό ισχύ των μπαταριών και την κατάσταση της ενέργειας τους εμφανίζονται στα σχήματα 5.44 και 5.45 αντίστοιχα.

Μέσω των γραφικών φαίνεται ότι η ισχύς τους είναι ανάλογη της μέγιστης τιμής, οπότε η χρήση είναι δίκαιη, όπως ορίζει το κριτήριο δικαιοσύνης 2.5.1.

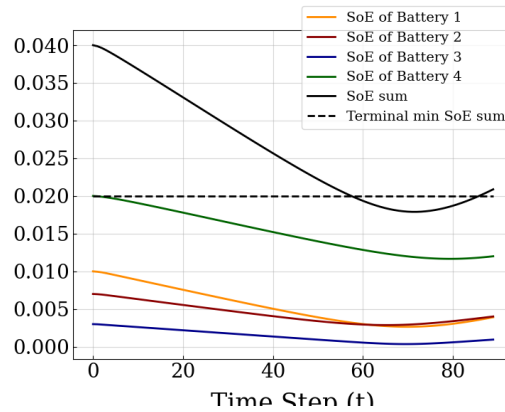
Σύγκριση αποτελεσμάτων

Σχετικά με την αντικειμενική αξιολόγηση των μεθόδων, ακολουθείται η ίδια μεθοδολογία, δηλαδή εκτελούνται 5 διαφορετικά παραδείγματα εκπαίδευσης και υπολογίζονται οι μετρικές που ορίστηκαν αρχικά. Τα αποτελέσματα που παρουσιάζονται αφορούν τον μέσο όρο αυτών των μετρικών.

Αρχικά, αξιολογείται η επίδοση των μεθόδων σε σχέση με το κόστος δικτύου. Τα αποτε-



Σχήμα 5.44: Ενέργος ισχύς μπαταριών (Περίπτωση 3, Μέθοδος 7)



Σχήμα 5.45: Κατάσταση της Ενέργειας των μπαταριών (Περίπτωση 3, Μέθοδος 7)

λέσματα εμφανίζονται στον επόμενο πίνακα :

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	1.537	1.258	1.539	1.894
$(SoE(t), P(t-1), B(t-1), t)$	1.601	1.458	1.508	1.601

Πίνακας 5.10: Αποτελέσματα μεθόδων MADDPG για το κόστος (Περίπτωση 3)

Συγκρίνοντας τις δύο αναπαραστάσεις που εξετάστηκαν παρατηρείται ότι για την περίπτωση των 4 μπαταριών σε 2 μεθόδους η αναπαράσταση που χρησιμοποιεί την εκτίμηση του φορτίου δίνει καλύτερα αποτελέσματα, ενώ στις άλλες δύο υπερτερεί η αναπαράσταση με τη συνολική ισχύ στο PCC. Τα αποτελέσματα, όμως, είναι συγκρίσιμα, οπότε δεν υπάρχει κάποια φανερή υπεροχή για κάποια από τις δύο αναπαραστάσεις.

Σε σχέση με τη συνάρτηση ανταμοιβής που χρησιμοποιείται, φαίνεται ότι και σε αυτή την περίπτωση οι μέθοδοι που χρησιμοποιούν τροχιές δίνουν τα καλύτερα αποτελέσματα, χωρίς όμως οι υπόλοιπες να υστερούν σημαντικά, ώστε να κρίνονται ανεπαρκείς.

Έπειτα, αξιολογούνται οι μέθοδοι ως προς το κριτήριο δικαιοσύνης 2.5.1. Προς αυτή την κατεύθυνση παρουσιάζονται στο πίνακα που ακολουθεί η μέση τιμή της μετρικής $fairness_metric_1$ για κάθε μέθοδο :

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	0.757	0.304	0.379	0.274
$(SoE(t), P(t-1), B(t-1), t)$	0.729	0.290	0.325	0.395

Πίνακας 5.11: Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.1 (Περίπτωση 3)

Προκύπτει ότι οι μέθοδοι με τις τροχιές βάσει του αντίστοιχου κριτηρίου (3 και 4) και αυτές που εφαρμόζουν άμεσα το κριτήριο δικαιοσύνης στη συνάρτηση ανταμοιβής παρουσιάζουν σημαντικά καλύτερη επίδοση ως προς τη δίκαιη χρήση σε σχέση με τη βασική μέθοδο. Συγκρίνοντας τις 2 αναπαραστάσεις δεν φαίνεται να υπερτερεί η μία έναντι της άλλης, καθώς στις μισές υπερτερεί η μία και στις υπόλοιπες η άλλη.

Τέλος, εξετάζονται οι μέθοδοι ως προς τον στόχο για δικαιοσύνη σύμφωνα με το κριτήριο 2.5.2. Οι τιμές της μετρικής $fairness_metric_2$ φαίνονται στον ακόλουθο πίνακα:

Κατάσταση	Βασική	Τροχιές (2.5.1)	Τροχιές (2.5.2)	MSE
$(SoE(t), L(t-1), t)$	1.056	0.829	0.984	0.877
$(SoE(t), P(t-1), B(t-1), t)$	3.303	3.222	2.654	2.883

Πίνακας 5.12: Αποτελέσματα μεθόδων MADDPG για το κριτήριο δικαιοσύνης 2.5.2 (Περίπτωση 3)

Παρατηρείται ότι για την αναπαράσταση $(SoE(t), \tilde{L}(t), t)$ οι τιμές είναι κοντά στο 1 για όλες τις περιπτώσεις και σημαντικά καλύτερες από ότι για την εναλλακτική αναπαράσταση. Αυτό πιθανώς εξηγείται από τη σταθερότητα της αναπαράστασης. Αντίθετα, χρησιμοποιώντας την εναλλακτική αναπαράσταση τα αποτελέσματα είναι χειρότερα. Στην περίπτωση αυτή η μέθοδος 6 είναι αυτή που δίνει τα καλύτερα αποτελέσματα, όπως είναι αναμενόμενο, καθώς αυτή έχει σχεδιαστεί με αυτό τον στόχο.

Η μετρική $fairness_metric_2$ εξετάζει μόνο το αν συγκλίνουν οι τιμές των SoE των μπαταριών, χωρίς να λαμβάνει υπόψιν κάτι άλλο. Ένας σημαντικός στόχος του κριτηρίου δικαιοσύνης 2.5.2 είναι να μη φτάσει η κατάσταση της ενέργειας καμίας μπαταρίας σε οριακή τιμή. Από τις προσομοιώσεις παρατηρήθηκε ότι αυτό συνέβαινε σε κάθε παράδειγμα των μεθόδων 5 και 6, που στόχευαν σε αυτό το κριτήριο. Αντίθετα, στην περίπτωση των 4 μπαταριών για όλες τις άλλες μεθόδους σε ένα σημαντικό ποσοστό (τουλάχιστον σε 3 από τα 5 παραδείγματα) υπήρχε τουλάχιστον μία μπαταρία που άδειαζε στο διάστημα του ελέγχου.

Επίλογος

6.1 Συμπεράσματα

Στο πλαίσιο της παρούσας διπλωματικής σχεδιάστηκε ένα καταναεμημένο σύστημα ελέγχου πολλαπλών μπαταριών σε πραγματικό χρόνο, το οποίο στοχεύει στη δίκαιη χρήση των μπαταριών. Ο έλεγχος έγινε μέσω του αλγορίθμου MADDPG, ο οποίος είναι ένας αλγόριθμος ενισχυτικής μάθησης πολλαπλών πρακτόρων. Ο αλγόριθμος αυτός είναι model-free, online και off-policy και χρησιμοποιεί το σχήμα centralized training decentralized execution.

Σχετικά με τη δικαιοσύνη, που αποτελεί τον κύριο στόχο της παρούσας έρευνας προτάθηκαν δύο κριτήρια. Με σκοπό την εισαγωγή των κριτηρίων αυτών προτάθηκαν δύο μέθοδοι. Η πρώτη μέθοδος χρησιμοποιεί τροχιές, οι οποίες υπολογίζονται εξαρχής μέσω της επίλυσης ενός προβλήματος κυρτής βελτιστοποίησης χρησιμοποιώντας κάποια σενάρια για το μη ελεγχόμενο φορτίο. Η δεύτερη μέθοδος χρησιμοποιούσε απευθείας στη συνάρτηση ανταμοιβής το τετράγωνο της απόκλισης από τη δίκαιη χρήση ως ποινή. Αυτή η μέθοδος μπορεί να χρησιμοποιηθεί, όμως, μόνο για το ένα από τα δύο προτεινόμενα κριτήρια. Επίσης, εξετάστηκαν δύο πιθανές αναπαραστάσεις για την κατάσταση, όπου η μία χρησιμοποιεί το φορτίο των μη ελεγχόμενων στοιχείων, ενώ η άλλη τη συνολική ισχύ στο PCC.

Τα συστήματα που χρησιμοποιούν τροχιές βασίστηκαν στην εργασία [2]. Παρά τις σημαντικές διαφορές στα δύο προβλήματα, αυτές οι μέθοδοι πετύχαιναν εξαιρετικά αποτελέσματα, τόσο σε σχέση με τον στόχο δικτύου όσο και σχετικά με τη δικαιοσύνη. Η χρήση μάλιστα του τύπου υπολογισμού των βαρών από την αντίστοιχη εργασία, δίνει το πλεονέκτημα στη μέθοδο αυτή ότι λειτουργούσε η ίδια τιμή για την παράμετρο που εξισορροπεί τους δύο στόχους σε όλες τις περιπτώσεις που εξετάστηκαν.

Επίσης, εξετάστηκε η απευθείας εισαγωγή στη συνάρτηση ανταμοιβής του κριτηρίου δικαιοσύνης. Αυτή η μέθοδος βελτιώνει τις προηγούμενες μεθόδους όσον αφορά τη δίκαιη χρήση, ενώ σε ορισμένες περιπτώσεις λειτουργούσε καλύτερα από τις μεθόδους που βασίζονται σε τροχιές. Όμως, υπήρχαν περιπτώσεις που το κόστος δικτύου ήταν σημαντικά υψηλότερο από τις υπόλοιπες μεθόδους, ενώ το βάρος που εξισορροπούσε τους δύο στόχους έπρεπε να ρυθμιστεί για βέλτιστα αποτελέσματα σε διαφορετικές περιπτώσεις.

Τέλος, παρατηρήθηκε ότι η αναπαράσταση της κατάστασης που χρησιμοποιούσε τη συνολική ισχύ στο PCC οδηγούσε σε σχεδόν όλες τις περιπτώσεις για όλες τις μεθόδους σε καλύτερα αποτελέσματα, με την αναπαράσταση με αναμενόμενο φορτίο να έχει πλεονέκτημα σε κάποιες περιπτώσεις σχετικά με τη δίκαιη χρήση.

6.2 Αντικείμενα Μελλοντικής Έρευνας

Σε όλες τις περιπτώσεις που εξετάστηκαν, οι συνδυαστικοί περιορισμοί που ορίστηκαν αφορούσαν την τελική χρονική στιγμή. Επομένως, μια μελλοντική επέκταση αφορά την εξέταση των υπάρχοντων μεθόδων σε προβλήματα, όπου ορίζεται συνδυαστικός περιορισμός για ενδιάμεσες χρονικές στιγμές.

Επίσης, οι μέθοδοι που χρησιμοποιήθηκαν προϋποθέτουν την επικοινωνία με το PCC για κάθε χρονική στιγμή. Ένας εξιδανικευμένο αποκεντρωμένο σχήμα ελέγχου δεν θα απαιτούσε καμία επικοινωνία μεταξύ των διαφόρων στοιχείων του δικτύου. Ένα ανοιχτό πρόβλημα είναι η εύρεση μεθόδων για τον έλεγχο σε πραγματικό χρόνο που δεν θα απαιτείται επικοινωνία ή, τουλάχιστον, η επικοινωνία θα είναι περιορισμένη, υπό την έννοια ότι θα γίνεται για λιγότερες χρονικές στιγμές ή σε τοπικότερο επίπεδο.

Τέλος, στην εργασία εξετάστηκαν δίκτυα με μικρό πλήθος μπαταριών. Για μεγαλύτερα δίκτυα ο χρόνος εκπαίδευσης θα ήταν αυξημένος σημαντικά. Η μελλοντική έρευνα θα μπορούσε να αφορά την εξέταση των μεθόδων αυτών σε μεγαλύτερα δίκτυα και την αποδοτικότερη εκπαίδευση, ώστε να μπορεί να πραγματοποιηθεί σε εύλογο χρονικό διάστημα.

Παραρτήματα

Υπολογισμός Τροχιών και Δικαιοσύνη

Στο παράρτημα αυτό εξετάζεται αν η μέθοδος της [ενότητα 2.2](#) εξασφαλίζει δικαιοσύνη. Υπενθυμίζεται ότι το πρόβλημα βελτιστοποίησης που χρησιμοποιείται για τον υπολογισμό λύσεων είναι το ακόλουθο:

$$\min_{B_1(t), \dots, B_{N_B}(t)} \sum_{t=k}^{T-1} \left\{ \frac{1}{D} \sum_{d=1}^D [P_{PCC,d}(t) - P_{disp}(t)]^2 + w_c \cdot \sum_{i=1}^{N_B} |B_i(t)| \right\} \quad (A'.1)$$

$$\text{subject to } P_{PCC,d}(t) = \sum_{i=1}^{N_B} B_i(t) + L_d(t) \quad (A'.2)$$

$$SoE_i(t+1) = SoE_i(t) + B_i(t)\Delta t \quad (A'.3)$$

$$B_i^{min} \leq B_i(t) \leq B_i^{max} \quad (A'.4)$$

$$SoE_i^{min} \leq SoE_i(t+1) \leq SoE_i^{max} \quad (A'.5)$$

$$SoE_j^{min, \tau_j} \leq \sum_{i=1}^{N_B} a_{ji} SoE_i(\tau_j) \leq SoE_j^{max, \tau_j} \quad (A'.6)$$

$$\text{for } i \in \{1, 2, \dots, N_B\}, t \in \{k, \dots, T-1\}, j \in \{1, \dots, N_C\}$$

Η ανάλυση που ακολουθεί ότι ορίζεται μία διαμέριση $A_1, \dots, A_{N_p}$ του συνόλου $\{1, \dots, N_B\}$ και κάθε συνδυαστικός αφορά ακριβώς τις μπαταρίες του συνόλου A_p , για κάποιο $p \in \{1, \dots, N_p\}$. Επίσης, θεωρείται σταθερό το πλάνο διανομής.

A'.1 Αγνοώντας το battery cycling cost

Αρχικά, η ανάλυση αγνοεί τον όρο του battery cycling cost, δηλαδή ισχύει $w_c = 0$. Επιλέγεται ένα υποσύνολο της διαμέρισης, χωρίς βλάβη της γενικότητας το σύνολο A_1 και υποτίθεται ότι περιέχει $M_B \geq 2$ μπαταρίες. Η βασική παρατήρηση είναι ότι τόσο το κόστος δικτύου όσο και ο συνδυαστικός περιορισμός εξαρτάται από το συνολικό άθροισμα της ενεργού ισχύους την ίδια χρονική στιγμή και όχι από την ισχύ της κάθε μίας ξεχωριστά.

Πράγματι, το κόστος δικτύου για τη χρονική στιγμή t είναι:

$$cost(t) = \sum_{d=1}^D \left(L_d(t) + \sum_{i \in A_1} B_i(t) + \sum_{i \in \{1, \dots, N_B\} - A_1} B_i(t) \right)^2$$

Αντίστοιχα, για τους συνδυαστικούς περιορισμούς που εξαρτώνται από τις αποφάσεις της

χρονικής στιγμής t , δηλαδή για περιορισμό j με $t_j > t$. Ο περιορισμός θα έχει τη μορφή:

$$SoE_j^{min, t_j} \leq \sum_{i \in A_1} SoE_i(t_j) \leq SoE_j^{max, t_j}$$

Όμως, το SoE μπορεί, λόγω του περιορισμού (Α'.3), να γραφεί ως:

$$SoE_i(t) = SoE_i(k) + \sum_{\tau=k}^{t-1} B_i(\tau) \cdot \Delta t$$

Άρα, ο περιορισμός γράφεται:

$$\begin{aligned} SoE_j^{min, t_j} &\leq \sum_{i \in A_1} SoE_i(t_j) \leq SoE_j^{max, t_j} \Leftrightarrow \\ SoE_j^{min, t_j} &\leq \sum_{i \in A_1} SoE_i(k) + \sum_{\tau=k}^{t_j-1} B_i(\tau) \cdot \Delta t \leq SoE_j^{max, t_j} \Leftrightarrow \\ SoE_j^{min, t_j} &\leq \sum_{i \in A_1} \left[SoE_i(k) + \left(\sum_{\tau=k}^{t-1} B_i(\tau) + B_i(t) + \sum_{\tau=t+1}^{t_j-1} B_i(\tau) \right) \cdot \Delta t \right] \leq SoE_j^{max, t_j} \Leftrightarrow \\ SoE_j^{min, t_j} &\leq \sum_{i \in A_1} (B_i(t)) \cdot \Delta t + \sum_{i \in A_1} \left[SoE_i(k) + \left(\sum_{\tau=k}^{t-1} B_i(\tau) + \sum_{\tau=t+1}^{t_j-1} B_i(\tau) \right) \cdot \Delta t \right] \leq SoE_j^{max, t_j} \end{aligned}$$

Επομένως, όλες οι επιλογές των $B_i(t)$, $i \in A_1$ που συνεχίζουν να ικανοποιούν τους περιορισμούς της ίδιας μπαταρίας, δηλαδή τους (Α'.4), (Α'.5) είναι εφικτές λύσεις του προβλήματος και έχουν την ίδια τιμή αντικειμενικής συνάρτησης. Επομένως, λύσεις που διατηρούν σταθερό άθροισμα ενεργού ισχύος κάθε χρονική στιγμή με τη βέλτιστη, είναι και αυτές βέλτιστες. Ειδικά, αν το άθροισμα έχει μία τιμή κοντά στο μηδέν, μπορεί να υπάρχει βέλτιστη λύση, όπου μία μπαταρία επιλέγει τη μέγιστη ισχύ της και υπόλοιπες εκφορτίζονται, για να έρθει το άθροισμα κοντά στο 0. Αυτό είναι πιθανότατα εφικτό σε πολλές περιπτώσεις ενδιαφέροντος, οπότε η μέθοδος από μόνη της δεν αρκεί.

Α'.2 Θεωρώντας το battery cycling cost

Αν το βάρος w_c πάρει θετική τιμή, τότε στην αντικειμενική συνάρτηση λαμβάνεται υπόψιν και όρος του battery cycling cost. Σε αυτή την περίπτωση, γενικά λύσεις που έχουν ίδια συνολική ενεργό ισχύ δεν έχουν σίγουρα ίδια τιμή αντικειμενικής συνάρτησης. Όμως, αν επιλέγουν όλες οι μπαταρίες να μην εκφορτίσουν (δηλαδή $B_i(t) \geq 0, \forall i \in A_1$) και το άθροισμα της ενεργούς ισχύος είναι σταθερό, τότε η τιμή είναι ίδια. Πράγματι, ο πρώτος όρος παραμένει σταθερός, όπως στην προηγούμενη περίπτωση, ενώ το battery cycling cost είναι:

$$\begin{aligned} battery_cycling_cost(t) &= \sum_{i=1}^{N_B} |B_i(t)| = \sum_{i \in A_1} |B_i(t)| + \sum_{i \in \{1, \dots, N_B\} - A_1} |B_i(t)| \\ &= \sum_{i \in A_1} B_i(t) + \sum_{i \in \{1, \dots, N_B\} - A_1} |B_i(t)| \end{aligned}$$

Αντίστοιχα, αν για κάθε μπαταρία ισχύει ότι δε φορτίζει (δηλαδή $B_i(t) \leq 0$), τότε :

$$\begin{aligned} \text{battery_cycling_cost}(t) &= \sum_{i=1}^{N_B} |B_i(t)| = \sum_{i \in A_1} |B_i(t)| + \sum_{i \in \{1, \dots, N_B\} - A_1} |B_i(t)| \\ &= - \sum_{i \in A_1} B_i(t) + \sum_{i \in \{1, \dots, N_B\} - A_1} |B_i(t)| \end{aligned}$$

Επομένως, αρκεί να είναι σταθερό το άθροισμα των ενεργών ισχύων και το γινόμενο δύο οποιωνδήποτε ισχύων του συνόλου να μην είναι αρνητικό, για να έχει η αντικειμενική συνάρτηση την ίδια τιμή. Επίσης, εφόσον το άθροισμα των ενεργών ισχύων είναι σταθερό, τότε οι συνδυαστικοί περιορισμοί συνεχίζουν να ικανοποιούνται. Επομένως, μπορεί να υπάρχουν εφικτές λύσεις που κάποιες μπαταρίες δεν χρησιμοποιούνται και αξιοποιούνται μόνο οι υπόλοιπες για τους στόχους του δικτύου. Άρα, το battery cycling cost μπορεί να περιορίσει το σύνολο των βέλτιστων λύσεων σε λύσεις που είτε φορτίζουν είτε εκφορτίζουν όλες οι μπαταρίες, αλλά δεν είναι αρκετό για να εξασφαλιστεί η δίκαιη χρήση.

A.3 Απόδειξη ικανοποίησης των συνδυαστικών περιορισμών από τον αλγόριθμο 4.1

Η απόδειξη βασίζεται στο ότι για κάθε χρονική στιγμή θα ισχύει $\sum_{i=1}^{N_B} B_i(t) = \sum_{i=1}^{N_B} B_i^{OPT}(t)$, όπου η $B_i(t)$ η ενεργός ισχύς της μπαταρίας i τη χρονική στιγμή t , όπως υπολογίζεται από τον αλγόριθμο 4.1 και $B_i^{OPT}(t)$ η ενεργός λύση που προέκυψε από επίλυση του προβλήματος βελτιστοποίησης. Εφόσον ισχύει αυτό για τη χρονική στιγμή t_j στην οποία ορίζεται συνδυαστικός περιορισμός για όλες τις μπαταρίες, θα είναι:

$$\begin{aligned} \sum_{i=1}^{N_B} \text{SoE}_i(t_j) &= \sum_{i=1}^{N_B} \text{SoE}_i(k) + \sum_{\tau=k}^{t_j-1} \sum_{i=1}^{N_B} B_i(t) \\ &= \sum_{i=1}^{N_B} \text{SoE}_i(k) + \sum_{\tau=k}^{t_j-1} \sum_{i=1}^{N_B} B_i^{OPT}(t) \\ &= \sum_{i=1}^{N_B} \text{SoE}_i^{OPT}(t_j) \end{aligned}$$

Όμως ο συνδυαστικός περιορισμός έχει οριστεί στο πρόβλημα βελτιστοποίησης, οπότε τον ικανοποιεί η βέλτιστη λύση, άρα και η λύση που προκύπτει από τον αλγόριθμο 4.1. Μένει ναδειχθεί ότι πράγματι ισχύει η σχέση $\sum_{i=1}^{N_B} B_i(t) = \sum_{i=1}^{N_B} B_i^{OPT}(t)$.

Η απόδειξη θα γίνει για την περίπτωση όπου $P_{ref} < 0$, καθώς η απόδειξη για την περίπτωση $P_{ref} > 0$ είναι ανάλογη. Με P'_{ref} θα συμβολίζεται η τιμή του P_{ref} μετά την ανανέωση στο τρέχον βήμα του αλγορίθμου, ενώ με $B'_i(t)$ η ανανεωμένη τιμή για την ενεργό ισχύ της μπαταρίας i στο βήμα του αλγορίθμου που εξετάζεται.

Αρχικά, ισχύει ότι $P'_{ref} \leq 0$, σε κάθε επανάληψη του βρόχου. Αυτό προκύπτει, καθώς:

$$\begin{aligned}
 P'_{ref} &= \sum_{i=1}^{N_B} B_i^{OPT} - \sum_{i=1}^{N_B} B'_i(t) \\
 &= P_{ref} + \sum_{i=1}^{N_B} B_i(t) - \sum_{i=1}^{N_B} B'_i(t) \\
 &\leq P_{ref} + \sum_{i=1}^{N_B} B_i(t) - \sum_{i=1}^{N_B} (B_i(t) + d_i \cdot \frac{SoE_i(t) - SoE_i^{min}}{\sum_{j=1}^{N_B} d_i \cdot (SoE_j(t) - SoE_j^{min})} \cdot P_{ref}) \\
 &= 0
 \end{aligned}$$

Επομένως, για να τερματίσει ο αλγόριθμος πρέπει να είναι $P_{ref} = 0$. Πράγματι, αν σε ένα βήμα του βρόχου ισχύει για κάθε μπαταρία με $d_i \neq 0$ ότι $B'_i(t) > B_i^{min}$, τότε θα είναι:

$$\begin{aligned}
 P'_{ref} &= \sum_{i=1}^{N_B} B_i^{OPT}(t) - \sum_{i:d_i \neq 0} (B_i(t) + \frac{SoE_i(t) - SoE_i^{min}}{\sum_{j=1}^{N_B} d_i \cdot (SoE_j(t) - SoE_j^{min})} \cdot P_{ref}) - \sum_{i:d_i=0} B_i(t) \\
 &= P_{ref} + \sum_{i=1}^{N_B} B_i(t) - \sum_{i=1}^{N_B} B_i(t) - \sum_{i=1}^{N_B} d_i \cdot \frac{SoE_i(t) - SoE_i^{min}}{\sum_{j=1}^{N_B} (d_i \cdot (SoE_j(t) - SoE_j^{min}))} \cdot P_{ref}) \\
 &= 0
 \end{aligned}$$

Αν δεν γίνεται ισχύει η παραπάνω σχέση, τότε στο βήμα αυτό τουλάχιστον μία μπαταρία παίρνει την τιμή $B_i(t) = B_i^{min}$ και γίνεται $d_i = 0$. Επομένως, αν δεν έχει τερματίσει ο αλγόριθμος μετά από N_B επαναλήψεις, τότε θα είναι για κάθε μπαταρία $d_i = 0$ και $B_i(t) = B_i^{min}$. Οπότε θα είναι:

$$\begin{aligned}
 P'_{ref} &= \sum_{i=1}^{N_B} B_i^{OPT}(t) - \sum_{i=1}^{N_B} B_i(t) \\
 &= \sum_{i=1}^{N_B} B_i^{OPT}(t) - \sum_{i=1}^{N_B} B_i^{min} \\
 &\geq 0
 \end{aligned}$$

Αφού είναι $B_i^{OPT}(t) \geq B_i^{min}$, λόγω των περιορισμών του προβλήματος βελτιστοποίησης. Άρα θα είναι $P'_{ref} = 0$. Επομένως, σε κάθε περίπτωση ισχύει ότι $P'_{ref} = 0$, δηλαδή $\sum_{i=1}^{N_B} B_i(t) = \sum_{i=1}^{N_B} B_i^{OPT}(t)$, το οποίο είναι το ζητούμενο.

Υπολογισμός Βαρών

Β'.1 Περιγραφή Άνω Φράγματος

Ως άνω φράγμα για την απόκλιση της βέλτιστης λύσης $P_{opt}(t)$ από το πλάνο διανομής $P_{disp}(t)$, ορίστηκε η συνάρτηση $f_u(w_m)$, όπου w_m το βάρος για το οποίο προκύπτει η λύση $P_{opt}(t)$. Υπενθυμίζεται ότι η συνάρτηση $f_u(w_m)$ ορίζεται ως εξής:

$$K_t(w_m, d) = \frac{1}{w_m J'(P_{disp}(t) + d)} = \frac{1}{2 \cdot w_m \cdot d} \quad (\text{B'.1})$$

$$h(K) = \begin{cases} |P_{disp}(t) - P_{inf}(t)| & , K \leq 1 \\ |P_{disp}(t) - P_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}| & , K > 1 \end{cases} \quad (\text{B'.2})$$

$$g(w_m, d) = \max\{d, h(K_t(w_m, d))\} \quad (\text{B'.3})$$

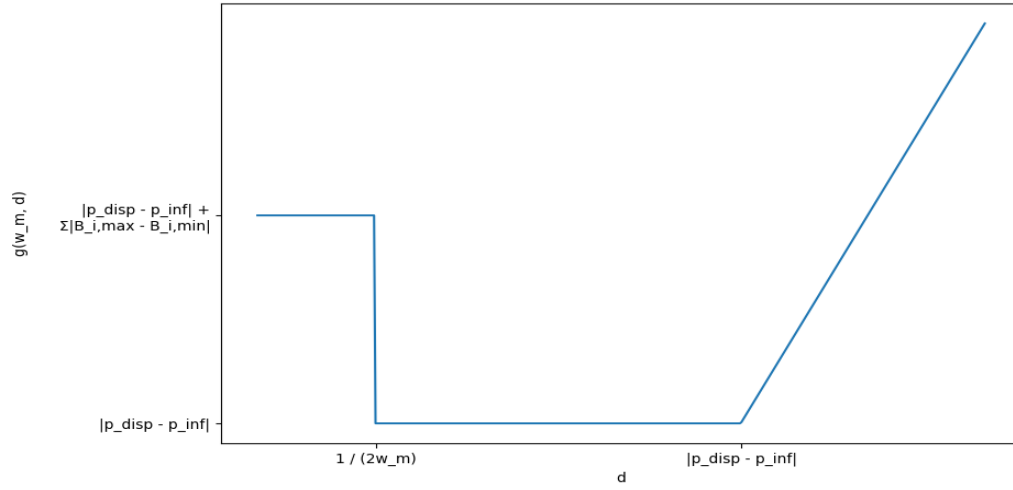
$$f(w_m) = \min_{d>0} g(w_m, d) \quad (\text{B'.4})$$

Ουσιαστικά, όπως θα γίνει φανερό στην απόδειξη ακολούθως, η συνάρτηση $g(w_m, d)$ είναι για κάθε d άνω φράγμα για την απόσταση $|P_{opt}(t) - P_{disp}(t)|$ και η συνάρτηση $f_u(w_m)$ αποτελεί το βέλτιστο άνω φράγμα.

Η συνάρτηση για συγκεκριμένο d έχει 3 πιθανές τιμές. Η περίπτωση $K > 1$, δίνει μία τιμή, που είναι άνω φράγμα για οποιαδήποτε απόκλιση $|P_{PCC}(t) - P_{disp}(t)|$, για εφικτό $P_{PCC}(t)$. Υπάρχει σαν προσασία, ώστε να εξακολουθεί η συνάρτηση να είναι άνω φράγμα για πολύ μικρά βάρη και μικρές αποκλίσεις, αφού η τιμή αυτή προκύπτει αν $K > 1 \Leftrightarrow 2 \cdot w_m \cdot d < 1$. Αν η τιμή του d είναι μικρότερη από $|P_{disp}(t) - P_{inf}(t)|$ και ισχύει $K \leq 1 \Leftrightarrow 2 \cdot w_m \cdot d \geq 1$, τότε η συνάρτηση παίρνει την τιμή $|P_{disp}(t) - P_{inf}(t)|$. Αυτό δηλώνει ότι αν η βέλτιστη λύση απέχει περισσότερο από d από το πλάνο διανομής, δηλαδή $|P_{opt}(t) - P_{disp}(t)|$ και για το d αυτό ισχύει $d \leq \frac{1}{2 \cdot w_m}$, τότε θα είναι $P_{opt}(t) = P_{inf}(t)$. Τέλος, για περιπτώσεις, οι περιπτώσεις όπου είναι $|P_{opt}(t) - P_{disp}(t)| \geq d$, καλύπτονται από τη χρήση του μεγιστού στη συνάρτηση $g(w_m, d)$.

Από την άλλη αν θεωρηθεί το βάρος w_m ως σταθερό, τότε η συνάρτηση $g_w(d) = g(w_m, d)$ έχει τρεις πιθανές μορφές.

Η πρώτη παρατηρείται για μεγάλα βάρη. Συγκεκριμένα, αν ισχύει ότι $w_m > \frac{1}{2 \cdot |P_{disp}(t) - P_{inf}(t)|}$, τότε παρατηρούνται και οι δύο δυνατές τιμές της συνάρτησης $h(K_t(w_m, d))$ και έχει τη μορφή που φαίνεται στην ακόλουθη γραφική παράσταση.



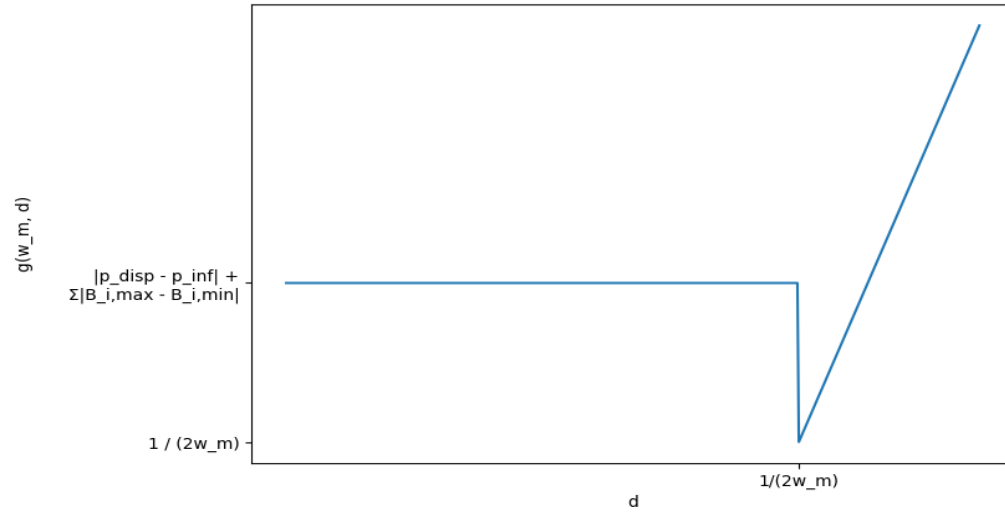
Σχήμα Β'.1: Η συνάρτηση $g(w_m, d)$: Πρώτη περίπτωση

Τότε, η συνάρτηση $f_u(w_m)$ θα έχει την τιμή $|p_{disp}(t) - p_{inf}(t)|$.

Μειώνοντας την τιμή του βάρους, ώστε να ισχύει:

$$\frac{1}{2 \cdot (|p_{disp}(t) - p_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}|)} < w_m < \frac{1}{2 \cdot |p_{disp}(t) - p_{inf}(t)|}$$

Η $g(w_m, d)$ δεν παίρνει την τιμή $|p_{disp}(t) - p_{inf}(t)|$, επομένως έχει την ακόλουθη μορφή:



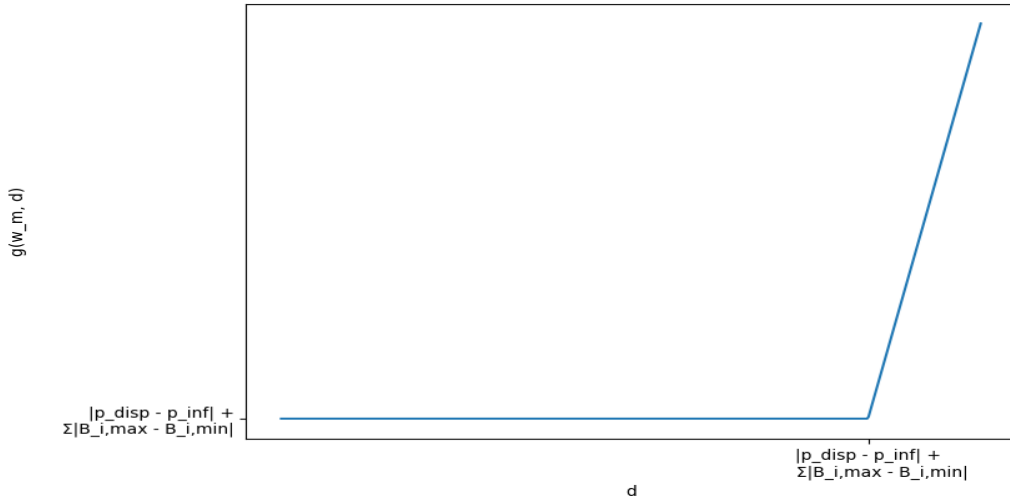
Σχήμα Β'.2: Η συνάρτηση $g(w_m, d)$: Δεύτερη περίπτωση

Επομένως, η τιμή της συνάρτησης $f_u(w_m)$ είναι ίση με $\frac{1}{2 \cdot w_m}$.

Τέλος, για πολύ μικρά βάρη, που ικανοποιούν δηλαδή τη σχέση:

$$w_m < \frac{1}{2 \cdot (|p_{disp}(t) - p_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}|)}$$

Παρατηρείται η ακόλουθη μορφή για τη συνάρτηση:


 Σχήμα Β'.3: Η συνάρτηση $g(w_m, d)$: Τρίτη περίπτωση

Τότε, η συνάρτηση $f_u(w_m)$ παίρνει την τιμή $|p_{disp}(t) - p_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}|$ και είναι σίγουρα άνω φράγμα για την τιμή $|P^{opt}(t) - P^{inf}(t)|$.

Β'.2 Απόδειξη γενικότερης πρότασης

Στην εργασία [2] χρησιμοποιούνταν η υπόθεση $a_p \leq 1$. Επειδή, στο πρόβλημα που πραγματεύεται η παρούσα εργασία οι τροχιές δεν είναι μέρος του αρχικού προβλήματος, αλλά χρησιμοποιούνται βοηθητικά, υπολογίζονται χρησιμοποιώντας σενάρια, τα οποία επιλέγονται από την ίδια κατανομή με τα σενάρια που παρουσιάζονται κατά το χρόνο εκτέλεσης. Γι' αυτό και αναμένεται οι τροχιές να είναι καλές λύσεις. Γι' αυτό και επιτρέπεται να δοθεί μεγαλύτερη τιμή στο a_p , ώστε να δίνεται μεγαλύτερη βαρύτητα στις τροχιές.

Βάσει της απόδειξης που παρουσιάζεται στο [2] και χρησιμοποιώντας απλοποιήσεις που παρουσιάζονται λόγω των υποθέσεων της παρούσας εργασίας, αποδεικνύεται ότι για κάποιο σταθερό $a_p \in [0, +\infty)$ η ακόλουθη πρόταση.

Λήμμα 2.2. *Αν η συνάρτηση για το grid objective $J(P_{PCC}(t))$ είναι αυστηρά κυρτή ως προς $P_{PCC}(t)$, η $J(P_{PCC}(t) - P_{disp}(t))$ άρτια, η ενεργός ισχύς της μπαταρίας έχει μόνο περιορισμούς της μορφής $B_i(t) \in [B_i^{min}, B_i^{max}]$ και ισχύει ότι $\left| \frac{B_i^{inf}(t) - B_i^*(SoE_i(t), t)}{B_i^{max}} \right| \leq \frac{2}{a_p}$, για κάθε μπαταρία i . Τότε η συνάρτηση $f_u(w_m)$ είναι άνω φράγμα της απόστασης $|P_{opt}(t) - P_{inf}(t)|$.*

Απόδειξη. Η συνάρτηση $f_u(w_m(t))$ ορίζεται ως $\min_{d>0} g(w_m(t), d)$, οπότε είναι αρκετό να δείχθει ότι $|P_{opt}(t) - P_{inf}(t)| \leq g(w_m(t), d)$, για κάθε $d > 0$. Υπάρχουν τρεις περιπτώσεις για την τιμή του d .

1η Περίπτωση: $d \geq |P_{opt}(t) - P_{inf}(t)|$

Ισχύει, γιατί:

$$\begin{aligned} g(w_m, d) &= \max\{d, h(K_t(w_m, d))\} \\ &\geq d \\ &\geq |P_{opt}(t) - P_{inf}(t)| \end{aligned}$$

2η Περίπτωση: $d < |P_{opt}(t) - P_{inf}(t)|$ και $d < \frac{1}{2 \cdot w_m}$

Επειδή, είναι $d < \frac{1}{2 \cdot w_m}$, ισχύει ότι $K_t(w_m, d) > 1$,

άρα $h(K_t(w_m, d)) = |p_{disp}(t) - p_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}|$.

Επομένως, η $g(w_m, d)$ είναι άνω φράγμα, γιατί:

$$\begin{aligned} g(w_m, d) &= \max\{d, h(K_t(w_m, d))\} \\ &\geq h(K_t(w_m, d)) \\ &= |p_{disp}(t) - p_{inf}(t)| + \sum_{i=1}^{N_B} |B_i^{max} - B_i^{min}| \\ &\geq |p_{disp}(t) - p_{inf}(t)| + |p_{inf}(t) - p_{opt}(t)| \\ &\geq |P_{opt}(t) - P_{inf}(t)| \end{aligned}$$

3η Περίπτωση: $d < |P_{opt}(t) - P_{inf}(t)|$ και $d > \frac{1}{2 \cdot w_m}$

Για κάποιο $d > 0$, είναι:

$$\begin{aligned} |P_{opt}(t) - P_{disp}(t)| &> d \Leftrightarrow \\ P_{opt}(t) - P_{disp}(t) &> d \text{ or } P_{opt}(t) - P_{disp}(t) < -d \\ P_{opt}(t) &> P_{disp}(t) + d \text{ or } P_{opt}(t) < P_{disp}(t) - d \end{aligned}$$

Όμως, η $J(P_{PCC}(t))$ είναι αυστηρά κυρτή, η παράγωγός της $J'(P_{PCC}(t))$ είναι γνησίως αύξουσα, οπότε προκύπτει ότι:

$$J'(P_{opt}(t)) > J'(P_{disp}(t) + d) \text{ or } J'(P_{opt}(t)) < J'(P_{disp}(t) - d)$$

Επειδή, η συνάρτηση $J(P_{PCC}(t) + P_{disp}(t))$ είναι άρτια, η συνάρτηση

$J'(P_{PCC}(t) + P_{disp}(t))$ είναι περιττή, άρα είναι $J'(P_{disp}(t) + d) = J'(P_{disp}(t) - d)$, για κάθε d .

Επομένως, θα είναι:

$$|J'(P_{opt}(t))| > |J'(P_{disp}(t) \pm d)|$$

Όμως, $w_m = \frac{1}{K_t(w_m, d) \cdot J'(P_{disp}(t) + d)}$, άρα για κάθε μπαταρία i , ισχύει:

$$\begin{aligned} w_m \cdot \left| \frac{\partial J(P_{opt})}{\partial B_i(t)} \right| &= \frac{1}{K_t(w_m, d) \cdot J'(P_{disp}(t) + d)} \cdot \left| \frac{\partial J(P_{opt})}{\partial B_i(t)} \right| \\ &= \left| \frac{J'(P_{opt}(t))}{J'(P_{disp}(t))} \right| \cdot \frac{1}{K_t(w_m, d)} \\ &> 1 \end{aligned} \tag{B'.5}$$

Ακόμη, επειδή ισχύει $\left| \frac{B_i^{inf}(t) - B_i^*(SoE_i(t), t)}{B_i^{max}} \right| \leq \frac{2}{a_p}$ θα είναι $\left| \frac{B_i^{opt}(t) - B_i^*(SoE_i(t), t)}{B_i^{max}} \right| \leq \frac{2}{a_p}$.

Άρα, προκύπτει ότι:

$$a_p \cdot \left| \frac{\partial C_i(t)}{\partial B_i(t)} \right| \leq 1 \tag{B'.6}$$

Από (Β'.5) και (Β'.6) προκύπτει ότι:

$$\frac{\partial C_t}{\partial B_i(t)} \neq 0$$

Αν είναι $w_m \cdot \frac{\partial J(P_{opt})}{\partial B_i(t)} > 1$, τότε είναι $\frac{\partial C_t}{\partial B_i(t)} > 0$, για κάθε i , οπότε είναι $B_i(t) = B_i^{max}$. Για να είναι προκύψει όμως η λύση αυτή, θα πρέπει να ισχύει $P_{disp}(t) > P_{inf}$, οπότε είναι:

$$\begin{aligned} P_{opt}(t) &= \sum_{i=1}^{N_B} B_i(t) + \tilde{L}(t) \\ &= \sum_{i=1}^{N_B} B_i^{max} + \tilde{L}(t) \\ &= P_{inf}(t) \end{aligned}$$

Άρα, ισχύει η ανισότητα $|P_{opt}(t) - P_{disp}(t)| \leq |P_{inf}(t) - P_{disp}(t)|$ και κατ' επέκταση η ανισότητα $|P_{opt}(t) - P_{disp}(t)| \leq g(w_m, d)$. Ομοίως, αποδεικνύεται και για την περίπτωση $w_m \cdot \frac{\partial J(P_{opt})}{\partial B_i(t)} < 1$ ότι $|P_{opt}(t) - P_{disp}(t)| \leq g(w_m, d)$.

Επομένως, σε κάθε περίπτωση ισχύει $|P_{opt}(t) - P_{disp}(t)| \leq f_u(w_m)$. $\square$

Ουσιαστικά, επιτρέποντας μεγαλύτερες τιμές, για το a_p , το πρόβλημα βελτιστοποίησης δίνει την καλύτερη δυνατή λύση, αν κάθε μπαταρία επιτρέπεται να επιλέξει ενεργό ισχύ από το διάστημα $[max\{B_i^{min}, B_i^*(SoE_i(t), t) - \frac{2}{a_p} B_i^{max}, min\{B_i^{max}, B_i^*(SoE_i(t), t) + \frac{2}{a_p} B_i^{max}, \}\}]$. Αυτό, σε περιπτώσεις όπου οι τροχιές δίνουν σχετικά καλά αποτελέσματα δεν είναι σημαντικό πρόβλημα και μπορεί να είναι χρήσιμο για να επιτευχθεί η δίκαιη χρήση.

Β'.3 Μέθοδος 2.4 και Κριτήριο Δικαιοσύνης 2.5.1

Στην υποενότητα αυτή αποδεικνύεται η ακόλουθη πρόταση.

Λήμμα 2.3. *Αν οι τροχιές για το SoE έχουν υπολογιστεί χρησιμοποιώντας τον περιορισμό $\frac{B_i^{tr}(t)}{B_i^{max}} = \frac{B_j^{tr}(t)}{B_j^{max}}$, τότε για τη θύση του προβλήματος βελτιστοποίησης, το οποίο επιλύεται ανεξάρτητα κάθε χρονική στιγμή, ισχύει $\frac{B_i^{opt}(t)}{B_i^{max}} = \frac{B_j^{opt}(t)}{B_j^{max}}$, για κάθε χρονική στιγμή, εφόσον είναι $P_{PCC}(t) = L(t) + \sum_{i=1}^{N_B} B_i(t)$ και η προτίμηση υπολογίζεται για κάθε μπαταρία σύμφωνα με τη σχέση (2.14), για $\epsilon_i = B_i^{max} \cdot \Delta t$.*

Απόδειξη. Αρχικά, ισχύει ότι $\frac{B_i^*(SoE_i(k), k)}{B_i^{max}} = \frac{B_j^*(SoE_j(k), k)}{B_j^{max}}$, γιατί $\frac{B_i^{tr}(k)}{B_i^{max}} = \frac{B_j^{tr}(k)}{B_j^{max}}$. Αυτό ισχύει, αφού θα είναι $B_i^*(SoE_i(k), k) = B_i^{tr}(k)$, $\forall i$, γιατί οι τροχιές ικανοποιούν τους περιορισμούς, και είναι $SoE_i(k) = SoE_i^*(k)$, $\forall i$, όπου $SoE_i(t)$ συμβολίζεται η παρατηρούμενη τιμή του SoE της μπαταρίας i τη χρονική στιγμή t , ενώ $SoE_i^*(t)$ η αντίστοιχη τιμή σύμφωνα με την τροχιά. Επομένως, ικανοποιούνται οι συνθήκες, για να ισχύει η εξίσωση (2.12). Άρα, θα είναι

$$\begin{aligned} \frac{B_i^{opt}(k) - B_i^*(SoE_i(k), k)}{B_i^{max}} &= \frac{B_j^{opt}(k) - B_j^*(SoE_j(k), k)}{B_j^{max}} \Rightarrow \\ \frac{B_i^{opt}(k)}{B_i^{max}} &= \frac{B_j^{opt}(k)}{B_j^{max}} \end{aligned} \quad (\text{Β'.7})$$

Θα αποδειχθεί με χρήση επαγωγής ότι για κάθε χρονική στιγμή t ισχύει η σχέση $\frac{B_i^{opt}(t)}{B_i^{max}} = \frac{B_j^{opt}(t)}{B_j^{max}}$. Έχει ήδη αποδειχθεί ότι η σχέση ισχύει για $t = k$. Έστω ότι για κάθε χρονική στιγμή $\tau < t$ ισχύει ότι $\frac{B_i^{opt}(\tau)}{B_i^{max}} = \frac{B_j^{opt}(\tau)}{B_j^{max}}$. Ακόμη, ισχύει ότι:

$$\begin{aligned} \frac{SoE_i(t) - SoE_i^*(t+1)}{B_i^{max}} &= \frac{SoE_i(k) + \sum_{\tau=k}^{t-1} B_i(\tau) \cdot \Delta t - SoE_i(k) - \sum_{\tau=k}^t B_i^{tr}(\tau) \cdot \Delta t}{B_i^{max}} \\ &= \frac{SoE_j(k) + \sum_{\tau=k}^{t-1} B_j(\tau) \cdot \Delta t - SoE_j(k) - \sum_{\tau=k}^t B_j^{tr}(\tau) \cdot \Delta t}{B_j^{max}} \\ &= \frac{SoE_j(t) - SoE_j^*(t+1)}{B_j^{max}} \end{aligned}$$

Το οποίο, μαζί με τη σχέση (2.14) δίνουν ότι ισχύει $\frac{B_i^*(SoE_i(t),t)}{B_i^{max}} = \frac{B_j^*(SoE_j(t),t)}{B_j^{max}}$, καθώς είναι:

$$\frac{B_i^*(SoE_i(t),t)}{B_i^{max}} = \begin{cases} 1 & , \frac{SoE_i(t) - SoE_i^*(t+1)}{B_i^{max}} \leq -\Delta t \\ -\frac{1}{\Delta t} \cdot \frac{SoE_i(t) - SoE_i^*(t+1)}{B_i^{max}} & , \left| \frac{SoE_i(t) - SoE_i^*(t+1)}{B_i^{max}} \right| < \Delta t \\ -1 & , \frac{SoE_i(t) - SoE_i^*(t+1)}{B_i^{max}} \geq \Delta t \end{cases}$$

Άρα, ισχύει η εξίσωση (2.12) και ισχύει ότι:

$$\frac{B_i^{opt}(t)}{B_i^{max}} = \frac{B_j^{opt}(t)}{B_j^{max}}$$

Για κάθε χρονική στιγμή t , δηλαδή οι βέλτιστες λύσεις είναι και δίκαιες σύμφωνα με το κριτήριο 2.5.1. □

Παράμετροι

Μπαταρία	SoE(0)	SoE ^{min}	SoE ^{max}	B ^{min}	B ^{max}
Μπαταρία 1	0.010	0	0.0334	-0.048	0.048
Μπαταρία 2	0.007	0	0.0334	-0.028	0.028
Μπαταρία 3	0.003	0	0.0334	-0.015	0.015
Μπαταρία 4	0.020	0	0.0334	-0.048	0.048

Πίνακας Γ'.1: Χαρακτηριστικά μπαταριών (σε ανά μονάδα τιμές)

Παράμετρος	Σύμβολο	Τιμή
Ελάχιστο άθροισμα των SoE (Περίπτωση 1)	SoE_A^{min}	0.0100
Μέγιστο άθροισμα των SoE (Περίπτωση 1)	SoE_A^{max}	0.0668
Ελάχιστο άθροισμα των SoE (Περίπτωση 2)	SoE_B^{min}	0.0100
Μέγιστο άθροισμα των SoE (Περίπτωση 2)	SoE_B^{max}	0.0668
Ελάχιστο άθροισμα των SoE (Περίπτωση 3)	SoE_A^{min}	0.0200
Μέγιστο άθροισμα των SoE (Περίπτωση 3)	SoE_A^{max}	0.1336

Πίνακας Γ'.2: Παράμετροι προβλημάτων βελτιστοποίησης (σε ανά μονάδα τιμές)

Υπερπαράμετρος	Σύμβολο	Τιμή
Μέγεθος Replay Buffer	$\mathcal{D}$	1000000
Μέγεθος Terminal Replay Buffer	$\mathcal{D}_T$	1000000
Μέγεθος συνόλου εκπαίδευσης (Batch size)	M	200
Ποσοστό δειγμάτων από τον Terminal Replay Buffer	-	0.7
Ρυθμός μάθησης actor	-	0.00001
Ρυθμός μάθησης critic	-	0.0001
Συνάρτηση βελτιστοποίησης	-	Adam
Ρυθμός ενημέρωσης target networks	τ	0.001
Μέγιστη τυπική απόκλιση Gaussian Noise	-	0.01
Ελάχιστη τυπική απόκλιση Gaussian Noise	-	0.001

Πίνακας Γ'.3: Τιμές Υπερπαραμέτρων

Βιβλιογραφία

- [1] Moslem Uddin, Huadong Mo, Daoyi Dong, Sondoss Elsayah, Jianguo Zhu και Josep Guerrero. *Microgrids: A review, outstanding issues and future trends*. *Energy Strategy Reviews*, 49:101127, 2023.
- [2] Plouton Grammatikos, Mario Paolone και Jean Yves Boudec. *Design of cost functions for the real-time control of microgrids hosting distributed energy-storage systems*, 2023.
- [3] M.F. Roslan, M. A. Hannan, Pin Jern Ker, M. Mannan, Kashem Muttaqi και T M Indra Mahlia. *Microgrid control methods toward achieving sustainable energy management: A bibliometric analysis for future directions*. *Journal of Cleaner Production*, 348:131340, 2022.
- [4] Joel André, Eleni Stai, Ognjen Stanojev και Gabriela Hug. *Battery control with look-ahead constraints in distribution grids using reinforcement learning*. *Electric Power Systems Research*, 211:108551, 2022.
- [5] Eleni Stai, Ognjen Stanojev, Riccardode Nardis di Prata και Gabriela Hug. *Distributed Reinforcement Learning for Real-Time Batteries Control Using Lagrangian Decomposition*. σελίδες 1–6, 2022.
- [6] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel και Igor Mordatch. *Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments*. *CoRR*, abs/1706.02275, 2017.
- [7] Μ. Παπαδοστεφανάκη. *Καταναεμημένη Ενισχυτική Μάθηση για τη Βέλτιστη Διαχείριση Καταναεμημένων Μπαταριών σε Smart Grids*. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2024.
- [8] Flora Charbonnier, Thomas Morstyn και Malcolm D. Mcculloch. *Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility*. *ArXiv*, abs/2203.03417, 2022.
- [9] Sergio Rozada, Dimitra Apostolopoulou και Eduardo Alonso. *Load frequency control: A deep multi-agent reinforcement learning approach*. *2020 IEEE Power & Energy Society General Meeting (PESGM)*, σελίδες 1–5. IEEE, 2020.
- [10] Eleni Stai, Lorenzo Reyes-Chamorro, Fabrizio Sossan, Jean Yves Le Boudec και Mario Paolone. *Dispatching Stochastic Heterogeneous Resources Accounting for Grid and Battery Losses*. *IEEE Transactions on Smart Grid*, 9(6):6522–6539, 2018.

- [11] Lantao Xing, Yateendra Mishra, Yu Chu Tian, Gerard Ledwich, Chunjie Zhou, Wenli Du και Feng Qian. *Distributed State-of-Charge Balance Control With Event-Triggered Signal Transmissions for Multiple Energy Storage Systems in Smart Grid*. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(8):1601–1611, 2019.
- [12] S.P. Boyd και L. Vandenberghe. *Convex Optimization*. Αριθμός μ. 1 στο Berichte über verteilte messsysteme. Cambridge University Press, 2004.
- [13] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., USA, 1η έκδοση, 1997.
- [14] Stuart Russell. *Artificial intelligence: A modern approach*, 2010.
- [15] Svante Wold, Kim Esbensen και Paul Geladi. *Principal component analysis*. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.
- [16] Shai Shalev-Schwartz και Shai Ben-David. *Understanding Machine Learning: From Theory to Practice*. Cambridge University Press, 2014.
- [17] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho και Haifeng Chen. *Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection*. *International Conference on Learning Representations*, 2018.
- [18] JinHyun Joa, SangWon Bangb και GeunDuk Parka. *Implementation of a Recommendation System Using Association Rules and Collaborative Filtering*. *Procedia Computer Science*, 91:944–952, 2016. Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016).
- [19] Vladimir Yu. Kiselev, Tallulah S. Andrews και Martin Hemberg. *Challenges in unsupervised clustering of single-cell RNA-seq data*. *Nature Reviews Genetics*, 20:273 – 282, 2019.
- [20] Andrew G Barto Richard S Sutton. *Reinforcement learning: An Introduction*. MIT Press, 2η έκδοση, 2018. <http://www.incompleteideas.net/book/the-book.html>.
- [21] Benjamin Recht. *A Tour of Reinforcement Learning: The View from Continuous Control*, 2018.
- [22] Jonas Buchli, Farbod Farshidian, Alexander Winkler, Timothy Sandy και Markus Giffthaler. *Optimal and Learning Control for Autonomous Robots*. 2017.
- [23] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring,

- Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps και David Silver. *Grandmaster level in StarCraft II using multi-agent reinforcement learning*. *Nature*, 575:350 – 354, 2019.
- [24] Charu C Aggarwal. *Neural Networks and Deep Learning: A Textbook, 2nd edition*. Cham, Switzerland: Springer International Publishing, 2023.
- [25] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [26] David Silver. *Lectures on Reinforcement Learning*. URL: <https://www.davidsilver.uk/teaching/>, 2015. Ημερομηνία πρόσβασης: 21-08-2024.
- [27] Michael L. Littman. *Markov games as a framework for multi-agent reinforcement learning*. *Machine Learning Proceedings 1994* William W. Cohen και Haym Hirsh, επιμελητές, σελίδες 157–163. Morgan Kaufmann, San Francisco (CA), 1994.
- [28] OpenAI. *Spinning up*. <https://spinningup.openai.com/en/latest/>. Ημερομηνία πρόσβασης: 21-08-2024.
- [29] Csaba Szepesvári. *Algorithms for Reinforcement Learning*, τόμος 4. Morgan & Claypool Publishers, 2010.
- [30] David Silver, Guy Lever, Nicolas Manfred Otto Heess, Thomas Degris, Daan Wierstra και Martin A. Riedmiller. *Deterministic Policy Gradient Algorithms*. *International Conference on Machine Learning*, 2014.
- [31] TP Lillicrap. *Continuous control with deep reinforcement learning*. *arXiv preprint arXiv:1509.02971*, 2015.
- [32] Sven Gronauer και Klaus Diepold. *Multi-agent deep reinforcement learning: a survey*. *Artificial Intelligence Review*, 55:895 – 943, 2021.
- [33] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke και Travis E. Oliphant. *Array programming with NumPy*. *Nature*, 585(7825):357–362, 2020.
- [34] J. D. Hunter. *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering*, 9(3):90–95, 2007.
- [35] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat

- Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu και Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015. Software available from tensorflow.org.
- [36] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG και others. *Gymnasium: A Standard Interface for Reinforcement Learning Environments*. *arXiv preprint arXiv:2407.17032*, 2024.
- [37] *Sequential Least Squares Programming (SLSQP)*. URL: <https://docs.quantum.ibm.com/api/qiskit/0.36/qiskit.algorithms.optimizers.SLSQP>. Ημερομηνία Πρόσβασης: 10-10-2024.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

π.χ.	παραδείγματος χάριν
ΑΠΕ	Ανανεώσιμες Πηγές Ενέργειας
ΜΔΑ	Μαρκοβιανή Διαδικασία Απόφασης
PCC	Point of Common Coupling
MDP	Markov Decision Process
MLP	Multi-Layer Perceptron
CNN	Convolution Neural Network
MSE	Mean Squared error
DDPG	Deep Deterministic Policy Gradient
MARL	Multi-Agent Reinforcement Learning
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
SoE	State of Energy
DP	Dispatch Plan