Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Κατανομή πόρων σε Ασύρματα Δίκτυα Πολλαπλής Πρόσβασης Διαίρεσης Ρυθμού με χρήση Ενισχυτικής Μάθησης

Διπλωματικη Εργασια

του

ΓΕΩΡΓΙΟΥ Κ. ΚΑΨΑΛΗ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής

Κατανομή πόρων σε Ασύρματα Δίκτυα Πολλαπλής Πρόσβασης Διαίρεσης Ρυθμού με χρήση Ενισχυτικής Μάθησης

Διπλωματικη Εργασια

του

ΓΕΩΡΓΙΟΥ Κ. ΚΑΨΑΛΗ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11η Νοεμβρίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Συμεών Παπαβασιλείου	Ιωάννα Ρουσσάκη	Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π.	Επίκουρη Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής

(Υπογραφή)

.....

Γεωργιος Κ. Καψαλης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η τεχνολογία των δικτύων Πολλαπλής Πρόσβασης Διαίρεσης Ρυθμού (Rate Splitting Multiple Access,RSMA) προσφέρει μια νέα οπτική στον τρόπο που οι χρήστες αποκτούν πρόσβαση και διαχειρίζονται τους διαθέσιμους τηλεπικοινωνιακούς πόρους. Είναι ένα πολλά υποσχόμενο σχήμα πολλαπλής πρόσβασης το οποίο συνδυάζει τα οφέλη της Πολλαπλής Πρόσβασης Διαίρεσης Χώρου (Space Division Multiple Access,SDMA) και της Μη Ορθογωνικής Πολλαπλής Πρόσβασης (Non Orthogonal Multiple Access,NOMA). Σχετικά πρόσφατα το RSMA έχει αποδειχθεί ότι μπορεί να επιτύχει καλύτερη φασματική και ενεργειακή απόδοση σε σχέση με τα SDMA και NOMA σχήματα.

Παρόλο που το RSMA έχει εξαιρετικά πλεονεκτήματα, η βελτιστοποίηση της κατανομής της διαθέσιμης ισχύος στα προς μετάδοση μηνύματα σε σύστημα κατερχόμενης ζεύξης (downlink) είναι ένα πολύ δύσκολο πρόβλημα. Εδώ και δεκατίες η βελτιστοποίηση της κατανομής ισχύος ερευνάται κυρίως με αλγορίθμους που υποθέτουν πλήρη γνώση του τρόπου με τον οποίο μεταβάλλεται το τηλεπικοινωνιακό δίκτυο (model-based). Τα τελευταία όμως χρόνια έχουν αναδυθεί αρκετές τεχνικές που βασίζονται στην παρατήρηση του τηλεπικοινωνιακού δίκτύου χωρίς να απαιτείται η εκ των προτέρων γνώση αυτού (model-free). Από τις πιο δημοφιλείς τεχνικές που έκαναν την εμφάνισή τους τα τελευταία χρόνια είναι η Βαθιά Ενισχυτική Μάθηση η οποία θεωρείται ότι έχει αξιόλογες προοπτικές εφαρμογής στα ευφυή ασύρματα δίκτυα.

Στην παρούσα διπλωματική εργασία αναπτύχθηκε ένα νέο μοντέλο βελτιστοποίησης της κατανομής ισχύος σε ένα RSMA δίκτυο μονής κυψέλης (single-cell) και Μίας Εισόδου- Μίας Εξόδου (SISO) με σκοπό να μεγιστοποιηθεί η ενεργειακή του απόδοση. Το αλγοριθμικό αυτό μοντέλο βασίζεται σε τεχνικές Βαθιάς Ενισχυτικής Μάθησης και συγκεκριμένα στους αλγορίθμους Deep Q-Learning (DQL) και REINFORCE με μοντελοποίηση σε σύστημα πολλαπλών πρακτόρων (multi-agent).

Τα αποτελέσματα από την υλοποίηση του μοντέλου και την πραγματοποίηση μιας σειράς εκτενών προσομοιώσεων αναδεικνύουν τα πλεονεκτήματα που προσφέρει η εφρμογη τεχνικών Βαθιάς Μάθησης στη δυναμική διαχείριση πόρων σε RSMA δίκτυο πραγματικού χρόνου.

Λέξεις Κλειδιά

Πολλαπλή Πρόσβαση Διαίρεσης Ρυθμού, Βαθιά Ενισχυτική Μάθηση, Συστήματα Πολλαπλών Πρακτόρων, Δίκτυα $6^{\eta\varsigma}$ γενιάς, Ενεργειακή Απόδοση, Κατανομή πόρων

Abstract

The technology of Rate Splitting Multiple Access (RSMA) networks offers a new perspective on the way users access and manage available telecommunication resources. It is a promising multiple access scheme that combines the benefits of Space Division Multiple Access (SDMA) and Non Orthogonal Multiple Access (NOMA). Relatively recently RSMA has been shown to be able to achieve better spectral and energy efficiency than SDMA and NOMA schemes. Although RSMA has great advantages, optimizing the allocation of available power to the messages to be transmitted in a downlink system is a very difficult problem. For decades now, the optimization of power distribution has been researched mainly with algorithms that assume full knowledge of the way the telecommunication network changes (model based). However, in recent years several techniques have emerged that are based on the observation of the telecommunications network without requiring prior knowledge of it (model-free). Among the most popular techniques that have appeared is Deep Reinforcement Learning, which is considered to have significant application prospects in intelligent wireless networks. In this thesis, a new optimization model of power distribution in a single-cell and Single-Input-Single-Output (SISO) RSMA network was developed in order to maximize its energy efficiency. This algorithmic model is based on Deep Reinforcement Learning techniques, specifically the Deep Q-Learning (DQL) and REINFORCE algorithms with modeling in a multi-agent system. The results from the implementation of the model and the realization of a series of extensive simulations highlight the advantages offered by the application of Deep Learning techniques to dynamic resource management in a real-time RSMA network.

Keywords

Rate Splitting Multiple Access, Deep Reinforcement Learning, Multi-Agent Systems, 6G Networks, Energy Efficiency, Resource Allocation

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε υπό την επίβλεψη του καθηγητή του ΕΜΠ, κ. Συμεών Παπαβασιλείου τον οποίο θα ήθελα να ευχαριστήσω για την υποστήριξη του και τη συνεργασία του κατά τη διάρκεια της εκπόνησής της. Επιπλέον, θα ήθελα να ευχαριστήσω την υποψήφια διδάκτορα Μαρία Διαμαντή για την άψογη συνεργασία και τον ενδιαφέρον που επέδειξε ώστε να υπάρξει μία άρτια καθοδήγηση για την ολοκλήρωση της εργασίας. Επίσης ήθελα να ευχαριστήσω την επίκουρη καθηγήτρια του University of New Mexico (USA), κυρία Ειρήνη-Ελένη Τσιροπούλου για τις πολύ χρήσιμες συμβουλές που μου παρείχε.

Με αφορμή την ολοκλήρωση των σπουδών μου, θα ήθελα να δώσω ένα μεγάλο ευχαριστώ στην δίδυμη αδελφή και συμφοιτήτριά μου Ελένη-Ελπίδα με την οποία συμπορευτήκαμε αυτά τα πέντε χρόνια. Την ευχαριστώ πολύ για την ανιδιοτελή στήριξή της, ήτανε καθοριστική για την ολοκκήρωση των σπουδών μου. Επίσης ένα ιδιαίτερο ευχαριστώ στον πατέρα μου και συνάδελφο Κυριάκο που ήτανε πάντα δίπλα μου και με καθοδηγούσε σοφά σε οποιαδήποτε σημαντική μου απόφαση. Τέλος ευχαριστώ την οικογένεια μου και ιδιαίτερα τη μητέρα μου που με στήριζε και τους φίλους μου με τους οποίους μοιράστηκα τα φοιτητικά μου χρόνια στο Εθνικό Μετσόβιο Πολυτεχνείο.

Γεώργιος Κ. Καψάλης

Περιεχόμενα

Κατάλογος Σχημάτων

1 Εισαγωγή

1.1 Πρόλογος

Η τεχνική πολλαπλής πρόσβασης διαίρεσης ρυθμού (Rate Splitting Multiple Access,RSMA) είναι μια πολλά υποσχόμενη τεχνική πολλαπλής πρόσβασης στους διαθέσιμους ραδιοπόρους και αναμένεται να φέρει επανάσταση στον τρόπο με τον οποίο θα πραγματοποιούνται οι ασύρματες επικοινωνίες στα δίκτυα νεάς γενιάς, παρέχοντας υπηρεσίες επικοινωνίας υψηλού ρυθμού μετάδοσης, μικρής καθυστέρησης και ανθεκτικότητας στις μεταβολές του ασύρματου καναλιού. Η τεχνική αυτή προτείνει τη διαίρεση των προς μετάδοση μηνυμάτων σε ξεχωριστά μηνύματα και στη συνέχεια την πολυπλεξία αυτών των μηνυμάτων σε κοινές και ιδιωτικές ροές δεδομένων (common/private data streams). Σε αντίθεση με τις συμβατικές μεθόδους πρόσβασης στο δίκτυο, το RSMA δίνει τη δυνατότητα στο δίκτυο να διαχειρίζεται την διακαναλική παρεμβολή στον πομπό και στον δέκτη ταυτόχρονα, αυξάνοντας με αυτόν τον τρόπο τη φασματική απόδοση και την ενεργειακή του αποδοτικότητα, ενώ ακόμα θέτει νέα όρια στην ασφάλεια που μπορεί να επιτευχθεί στις ασύρματες μεταδόσεις δεδομένων.

Παρόλο που προκύπτουν σημαντικά πλεονεκτήματα, η βελτιστοποίηση της επίδοσης του RSMA είναι ένα ζήτημα που απασχολεί αρκετά την επιστημονική κοινότητα. Λόγω του μεγάλου πλήθους μηνυμάτων που προκύπτουν για μετάδοση ο πομπός πρέπει να αναθέτει προσεκτικά κατάλληλα επίπεδα ισχύος εκπομπής ώστε να ικανοποιούνται περιορισμοί που αφορούν στην ορθή λειτουργία του RSMA αλλά και περιορισμοί που αφορούν σε ορισμένους στόχους βελτιστοποίησης, όπως η ενεργειακή απόδοση. Επομένως, οι ερευνητές καλούνται να επιλύσουν ορισμένα προβλήματα βελτιστοποίησης πολλών περιορισμών και μεγάλης πολυπλοκότητας. Πολλές φορές οι κλασσικές μέθοδοι για ανάθεση ισχύος (power allocation) κρίνονται ανεπαρκείς. Για αυτό το λόγο, μέρος της επιστημονικής κοινότητας έχει στραφεί σε πιο εξελιγμένες τεχνικές βελτιστοποίησης. Μία σχετικά πρόσφατη κλάση αλγοριθμικών μεθόδων που βρίσκει εφαρμογή σε τηλεπικοινωνιακά προβλήματα είναι η Ενισχυτική Μάθηση. Η κλάση αυτή προσφέρει ένα αλγοριθμικό πλαίσιο για ακολουθιακά προβλήματα αποφάσεων και επιτρέπει μια προσαρμοστική ανάθεση ισχύος κάτω από ένα δυναμικό και αβέβαιο περιβάλλον επικοινωνίας. Ένα σημαντικό πλεονέκτημα έναντι των συμβατικών αλγορίθμων είναι ότι δεν απαιτείται εκ των προτέρων γνώση των πληροφοριών κατάστασης καναλιού (Channel State

Information-CSI).

Ειδικότερα η πρόοδος που έχει επιτευχθεί στον τομέα της Βαθιάς Μάθησης (Deep Learning) έχει ανοίξει νέους ορίζοντες σε πολλές περιοχές της μηχανικής μάθησης με εφαρμογές σε ένα μεγάλο φάσμα προβλημάτων αναζήτησης βέλτιστων λύσεων. Η πιο σημαντική συνεισφορά της βαθιάς μάθησης είναι ότι τα βαθιά νευρωνικά δίκτυα είναι ικανά να αναπαραστήσουν δεδομένα υψηλών διαστάσεων.

Στην παρούσα εργασία προτείνουμε ένα μοντέλο που δίνει τη δυνατότητα στον πομπό να προσαρμόζει την ισχύ κοινών και ιδιωτικών μηνυμάτων με βάση κάποια πολιτική που μαθαίνει παρατηρώντας τις μεταβολές των ασύρματων καναλιών. Ο κύριος στόχος της πολιτικής είναι να μεγιστοποιήσει κάποια επιθυμητή μετρική επίδοσης. Το μοντέλο που προτείνουμε στην παρούσα εργασία στοχεύει στη μεγιστοποίηση της ενεργειακής απόδοσης (Energy Efficiency,EE) του προς μελέτη ασύρματου τηλεπικοινωνιακού συστήματος.

1.2 Σχετική Έρευνα

Στα ασύρματα κυψελωτά δίκτυα νέας γενιάς αναμένεται ραγδαία αύξηση της πυκνότητας των χρηστών που είναι συνδεδεμένοι στο ίδιο δίκτυο λόγω του πολλαπλασιασμού των χρηστών που έχουν πρόσβαση σε τεχνολογίες Κινητής Τηλεφωνίας και στο Διαδίκτυο των Πραγμάτων (Internet of Things,IoT). Η συνέπεια είναι το ενεργειακό κόστος να αυξάνεται και να αποτελεί μείζονα απειλή για τη βιώσιμη ανάπτυξη των τηλεπικοινωνιακών εφαρμογών. Έχουν γίνει πολλές προσπάθειες να επιλυθεί το πρόβλημα της Ενεργειακής Βελτιστοποίησης και να επιτευχθεί ένα βέλτιστο σημείο ισορροπίας μεταξύ του σταθμισμένου αθροίσματος των ρυθμών μετάδοσης (Weighted Sum Rate,WSR) και της συνολικής ισχύος κατανάλωσης που απαιτείται για την επίτευξη του συγκεκριμένου WSR [Buz+16]. Στο [TTJ15], έχει μελετηθεί το πρόβλημα EE σε συστήματα Πολλαπλής Εισόδου-Μονής Εξόδου (Multiple Input- Single Output,MISO) υπό περιορισμούς μεγίστου αθροίσματος ισχύος και ελαχίστου επιτρεπτού Σηματοθορυβικού Λόγου (Signal Noise Ratio,SNR) στους χρήστες, με χρήση της Διαδοχικής Κυρτής Προσέγγισης (Successive Convex Approximation,SCA). Το ίδιο πρόβλημα βελτιστοποίησης επεκτείνεται σε δίκτυα πολλαπλών κελιών στο [Ter+17]. Μία διαφορετική αλγοριθμική προσέγγιση προτείνεται στο [Yan+18] με χρήση της Διαδοχικής Ψευδοκυρτής Προσέγγισης για επίτευξη βέλτιστου EE σε δίκτυα Πολλαπλής Εισόδου-Πολλαπλής Εξόδου (Multiple Input- Multiple Output,MIMO). Όλες οι εργασίες που προαναφέρθηκαν θεωρούν Πολλαπλή Πρόσβαση Διαίρεσης Χώρου (Space Division Multiple Access,SDMA). Σε αυτά ο δέκτης αποκωδικοποιεί το μήνυμα που προορίζεται για εκείνον αντιμετωπίζοντας την παρεμβολή από άλλα μηνύματα ως θόρυβο καναλιού.

Στην εργασία [Yan+18] ερευνάται η βελτιστοποίηση του EE για 2 χρήστες, σε Μη Ορθογωνικά σχήματα Πολλαπλής Πρόσβασης (Non Orthogonal Multiple Access,NOMA). Στα σχήματα NOMA, τα οποία είναι ευρέως χρησιμοποιούμενα στα σύγχρονα ασύρματα δίκτυα, οι δέκτες αποκωδικοποιούν την παρεμβολή στο πεδίο της ισχύος. Στο [Sun+15] , η βελτιστοποίηση επεκτείνεται σε συστήματα με περισσότερους από 2 χρήστες.

Το RSMA είναι ένα σχήμα πολλαπλής πρόσβασης που προτάθηκε ώστε να γεφυρώσει τα

πλεονεκτήματα των SDMA και NOMA σχημάτων. Η πρώτη εργασία που εξερεύνησε την EE στο RSMA είναι η [MCL18] και μελέτησε τα οφέλη του σε σχέση με τα συμβατικά σχήματα πολλαπλής πρόσβασης. Από τότε έχουν υπάρξει αρκετές ερευνητικές εργασίες στην περιοχή αυτή. Συγκεκριμένα η [ZMC21] εξετάζει το συμβιβασμό μεταξύ φασματικής και ενεργειακής απόδοσης με μεθόδους κυρτού προγραμματισμού για την επίλυση του προβλήματος βελτιστοποίησης που θέτει. Αυτό είναι και το βασικό ζήτημα της παρούσας εργασίας, να εξετάσουμε δηλαδή πως συμπεριφέρεται ένα τηλεπικοινωνιακό δίκτυο ως προς το WSR όταν καλούμαστε να μεγιστοποιήσουμε την EE και αντιστρόφως. Αντί να χρησιμοποιήσουμε πιο συμβατικές μεθόδους κυρτού προγραμματισμού, επιστρατεύουμε πιο σύγχρονες τεχνικές και συγκεκριμένα τεχνικές Μηχανικής Μάθησης (Machine Learning).

Η Μηχανική Μάθηση χωρίζεται σε δύο κύριες κατηγορίες. Η πρώτη κατηγορία ονομάζεται Επιβλεπόμενη Μάθηση (Supervised Learning) και χρησιμοποιείται σε προβλήματα όπου είναι γνωστές οι «ετικέτες» των δειγμάτων εκπαίδευσης και χρησιμοποιείται συνήθως σε προβλήματα ταξινόμησης. Τεχνικές επιβλεπόμενης μάθησης έχουν εφαρμοστεί για αναγνώριση του σημείου του αστερισμού που έχει χρησιμοποιηθεί για την ψηφιακή διαμόρφωση των προς μετάδοση παλμών από τον δέκτη [ZMC21] και για την ανίχνευση σήματος (signal detection) [YLJ18]. Ωστόσο, η επιβλεπόμενη μάθηση προσφέρει ένα πλαίσιο επίλυσης συγκεκριμένων προβλημάτων. Στα περισσότερα συστήματα δεν μπορούμε να γνωρίζουμε με βεβαιότητα τις «ετικέτες» των δειγμάτων, για αυτό και αναπτύχθηκε η δεύτερη κατηγορία μηχανικής μάθησης που ονομάζεται Ενισχυτική Μάθηση (Reinforcement Learning). Η Ενισχυτική Μάθηση εστιάζει στο πως ένας πράκτορας (agent) μπορεί να μαθαίνει να εκτελεί ενέργειες εντός ενός συγκεκριμένου περιβάλλοντος με σκοπό να μεγιστοποιήσει μία σωρευτική συνάρτηση ανταμοιβής (reward fuction). Η τομή της Ενισχυτικής Μάθησης με την περιοχή της Βαθιάς Μάθησης έχουν δημιουργήσει μία καινούργια ερευνητική περιοχή που ονομάζεται Βαθιά Ενισχυτική Μάθηση (Deep Reinforcement Learning,DRL), η οποία έχει απασχολήσει εκτενώς τους ερευνητές τα τελευταία χρόνια, καθώς έχει επιτύχει αξιόλογα αποτελέσματα σε αρκετές εφαρμογές [Li17] κάποιες εκ των οποίων είναι το παιχνίδι Go [al17] και το βιντεοπαιχνίδι Atari [al15].

Όσον αφορά εφαρμογές μηχανικής μάθησης σε πρόβληματα κατανομής ισχύος σε τηλεπικοινωνιακά συστήματα, η αρχή έγινε στις εργασίες [Sun+18] και [Lia+18] με στόχο την επίλυση του μη κυρτού προβλήματος βελτιστοποίησης του WSR σε δίκτυα πολλαπλών καναλιών που υποφέρουν από διαλείψεις. Η τεχνική (Deep Q-Learning,DQL) έχει χρησιμοποιηθεί εκτεταμένα για διαφορετικά τηλεπικοινωνιακά σενάρια ανάθεσης ισχύος όπως σε ετερογενή δίκτυα που αποτελούνται από έναν μακροσταθμό βάσης και πολλαπλές φεμτοκυψέλες (femtocells) [BN18], [Sim+11]. Επίσης, στο [Li+18] έχουν χρησιμοποιηθεί τεχνικές βαθιάς μάθησης για διαμοιρασμό φάσματος σε ευφυή ραδιοσυστήματα cognitive radio networks.

Τέλος, στις εργασίες [Men+20], [MCW19] και [NG19] έχει εξεταστεί η βελτιστοποίηση του WSR κυψελωτών δικτύων σε περιβάλλον πολλαπλών πρακτόρων, όπου οι πράκτορες λειτουργούν συνεργατικά για την επίτευξη του κοινού στόχου. Σε αυτές τις τρεις εργασίες έχει βασιστεί το αλγοριθμικό πλαίσιο που έχει προταθεί στην παρούσα εργασία προς βελτιστοποίηση του EE υπό συγκεκριμένους περιορισμούς που αφορούν στο σχήμα RSMA και στην μέγιστη

διαθέσιμη ισχύ εκπομπής. Εξ΄ όσων γνωρίζουμε, η παρούσα εργασία είναι η πρώτη που μελετάει το EE για RSMA με αλγοριθμικό μοντέλο επίλυσης που βασίζεται σε τεχνικές ενισχυτικής μάθησης και συγκεκριμένα σε ένα περιβάλλον πολλαπλών πρακτόρων.

1.3 Διάρθρωση της Διπλωματικής Εργασίας

Στο πρώτο κεφάλαιο, πραγματοποιείται μια συνοπτική αναφορά στο θέμα που απασχολεί την παρούσα εργασία, ακολουθούμενη από μια σύντομη ανασκόπηση σε σχετικές εργασίες. Στο δεύτερο κεφάλαιο, παρουσιάζεται το απαραίτητο θεωρητικό υπόβαθρο ώστε ο αναγνώστης να εισαχθεί στο περιβάλλον του προβλήματος που αφορά η εργασία. Συγκεκριμένα περιγράφονται σύντομα οι βασικές έννοιες των δικτύων πολλαπλής πρόσβασης (Multiple Access,MA) και κυρίως της Διαίρεσης Ρυθμού (Rate Splitting,RS) και της ενισχυτικής μάθησης που θα χρησιμοποιηθούν για την επίλυση του προβλήματος. Στο τρίτο κεφαλαίο, παρουσιάζεται το σύστημα επικοινωνίας υπό μελέτη, καθώς και η περιγραφή του προβλήματος βελτιστοποίησης. Στο τέταρτο κεφάλαιο, περιγράφεται αναλυτικά το αλγοριθμικό μοντέλο που επιλύει το πρόβλημα βελτιστοποίησης. Στο πέμπτο κεφάλαιο, παρατίθενται τα αριθμητικά αποτελέσματα που λαμβάνουμε από την εκτέλεση των υλοποιηθέντων προσομοιώσεων. Τα τελικά συμπεράσματα συνοψίζονται στο έκτο κεφάλαιο όπου και προτείνονται μελλοντικές επεκτάσεις.

2

Θεωρητικό Υπόβαθρο

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο περιγράφονται ορισμένες βασικές έννοιες αναφορικά με τις τεχνικές πολλαπλής πρόσβασης OFDMA (Orthogonal Frequency Division Multiple Access), SDMA και NOMA οι οποίες χρησιμοποιούνται στα σύγχρονα ασύρματα δίκτυα. Στη συνέχεια, περιγράφεται εκτενώς η τεχνική RSMA. Συγκεκριμένα, παρουσιάζεται η αρχιτεκτονική της καθώς και οι λειτουργίες που προσφέρει στα ασύρματα δίκτυα νέας γενιάς. Επιπλέον, παρέχεται μια σύντομη εισαγωγή στην Ενισχυτική Μάθηση (Reinforcement learning,RL) και αναλύονται σχετικές με αυτήν έννοιες που θα χρησιμοποιηθούν στην συνέχεια για την επίλυση του προβλήματος βελτιστοποίησης.

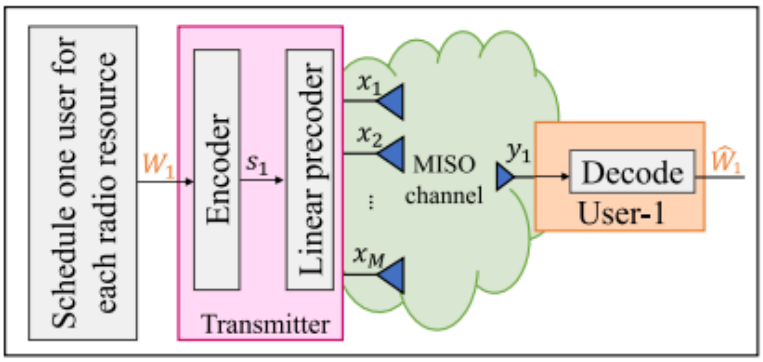
2.2 Multiple Access Techniques (MA)

Από τα ασύρματα δίκτυα πρώτη γενιάς 1G έως και τα σημερινά δίκτυα 5G, η εξέλιξη των τεχνικών πολλαπλής πρόσβασης αποτελεί αντικείμενο συνεχούς έρευνας και ανάπτυξης. Δεδομένου ότι το πλήθος των συσκευών που συνδέονται μέσω ασύρματης πρόσβασης στο Διαδίκτυο αυξάνεται συνεχώς υπάρχει ανάγκη για καλύτερη και αποδοτικότερη αξιοποίηση των διαθέσιμων πόρων. Για να μπορέσουν να ικανοποιηθούν οι αυξημένες απαιτήσεις τα ασύρματα δίκτυα $5^{ης}$ γενιάς θυσιάζουν την ορθογωνιότητα των ραδιοπόρων που ανατίθονται στους χρήστες προκειμένου να επιτευχθεί καλύτερη φασματική απόδοση.

2.2.1 Ορθογωνική Πολλαπλή Πρόσβαση (OMA)

Η βασική ιδέα της Ορθογωνικής Πολλαπλής Πρόσβασης είναι η ανάθεση ορθογώνιων μεταξύ τους πόρων στους χρήστες ώστε να αποφεύγεται η παρεμβολή με χρήση κατάλληλων φίλτρων. Στα ασύρματα δίκτυα πρώτης γενιάς χρησιμοποιούταν η τεχνική Πολλαπλής Πρόσβασης Διαίρεσης Συχνότητας (Frequency Division Multiple Access,FDMA), στην οποία το διαθέσιμο εύρος ζώνης χωρίζεται σε μη επικαλυπτόμενες μπάντες συχνοτήτων, όπου κάθε μπάντα ανατίθεται σε έναν μόνο χρήστη. Στα δίκτυα 2G χρησιμοποιήθηκε η τεχνική

Σχήμα 2.1: Απεικόνιση λειτουργίας της τεχνικής OMA για 2 χρήστες σε downlink μετάδοση [Mao+22].

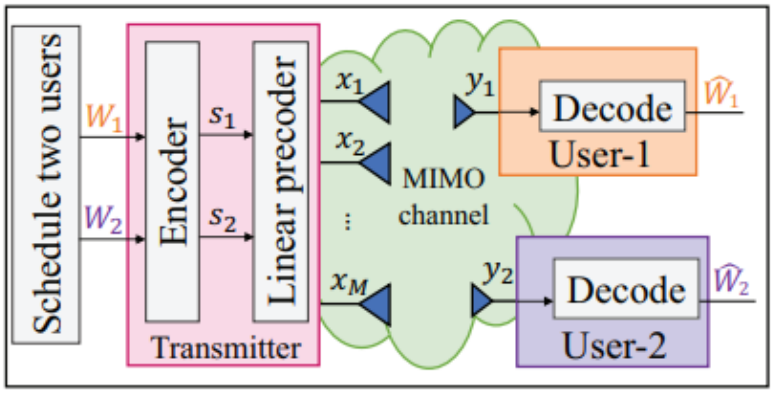


Πολλαπλής Πρόσβασης Διαίρεσης Χρόνου (Time Division Multiple Access,TDMA). Σύμφωνα με αυτήν την τεχνική, ο χρόνος χωρίζεται σε χρονικές σχισμές (time slots) οι οποίες δρομολογούνται στους χρήστες. Με αυτό τον τρόπο, πολλαπλοί χρήστες μπορούν να μοιράζονται την ίδια μπάντα συχνοτήτων. Στα δίκτυα 3G, επιστρατεύτηκε για πρώτη φορά η τεχνική Πολλαπλής Πρόσβασης Διαίρεσης Κώδικα (Code Division Multiple Access,CDMA), όπου πολλοί χρήστες μπορούν να στέλνουν μηνύματα ταυτόχρονα πάνω στο ίδιο κανάλι επικοινωνίας. Κατά την τεχνική αυτή, δρομολογούνται στους χρήστες ορθογώνιοι μεταξύ τους κώδικες που διευρύνουν τα προς μετάδοση στενής ζώνης σύμβολα στο πεδίο της συχνότητας. Εάν ο πομπός διαθέτει τους αντίστροφους κώδικες έχει τη δυνατότητα να ανακτήσει τα αρχικά σύμβολα στενής ζώνης. Στα πιο πρόσφατα δίκτυα 4G, οι ραδιοπόροι στο πεδίο της συχνότητας και του χρόνου διαιρούνται περαιτέρω σε υποφορείς και χρονικές σχισμές που φέρουν επιπρόσθετη πληροφορία. Η τεχνική αυτή είναι γνωστή ως Πολλαπλή Πρόσβαση Ορθογωνικής Διαίρεσης Συχνότητας (Orthogonal Frequency-Division Multiple Access,OFDMA). Οι τεχνικές που αναφέρθηκαν έχουν το πλεονέκτημα ότι αντιμετωπίζουν την παρεμβολή που προκαλείται μεταξύ των χρηστών με απλά φίλτρα στους δέκτες. Το μειονέκτημα, όμως, είναι ότι κάθε μπλοκ ραδιοπόρου (Resource Block) μπορεί να καταλαμβάνεται από έναν και μόνο χρήστη. Αυτό έχει σαν αποτέλεσμα ένα πρακτικό όριο στον αριθμό των χρηστών που μπορούν να εξυπηρετηθούν ταυτοχρόνως, ο οποίος είναι ανάλογος των διαθέσιμων ραδιοπόρων. Επίσης, είναι αναγκαία η ύπαρξη σύνθετων αλγορίθμων δρομολόγησης ώστε χρήστες με χαμηλές απαιτήσεις να μην δεσμεύουν μπλοκς που δεν χρειάζονται. Συμπερασματικά, οι ορθογωνικές τεχνικές πολλαπλής πρόσβασης έχουν χαμηλή φασματική απόδοση σε σύγχρονα συστήματα υψηλής κινητικότητας και χρηστών με μεταβαλλόμενες απαιτήσεις.

2.2.2 Πολλαπλή Πρόσβαση Διαίρεσης Χώρου (SDMA)

Η βασική ιδέα της τεχνικής SDMA είναι ότι οι σταθμοί βάσης (Base Stations,BS) διαθέτουν συστοιχίες κεραιών, δηλαδή σύνολα πολλαπλών συνδεδεμένων κεραιών που λειτουργούν ως ενιαίες κεραίες. Συνεπώς, μπορεί να αυξηθεί η χωρητικότητα του συστήματος σχη-

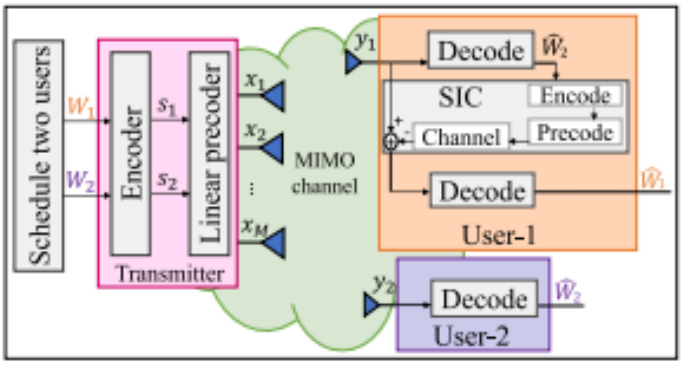
Σχήμα 2.2: Απεικόνιση λειτουργίας της τεχνικής SDMA για 2 χρήστες σε downlink μετάδοση [Mao+22].



ματίζοντας κατευθυνόμενες δέσμες ραδιοκυμάτων προς τις επιθυμητές κατευθύνσεις [Rap98]. Πολλοί χρήστες μπορούν να εξυπηρετηθούν ταυτόχρονα στο ίδιο κανάλι με υπέρθεση των αντίστοιχων κατευθυντικών ραδιοκυμάτων επιτρέποντας αύξηση στο δείκτη επαναχρησιμοποίησης συχνοτήτων σε κυψελωτά δίκτυα [ZO95]. Τα βασικά προβλήματα των συστημάτων SDMA είναι τα εξής δύο [DHM06],[Tan95a]: (i) Οι χρήστες που μοιράζονται την ίδια μπάντα συχνοτήτων πρέπει να βρίσκονται σε θέσεις που απέχουν γωνιακά μεταξύ τους κατά ένα ελάχιστο κατώφλι (threshold). Αυτό συμβαίνει διότι όταν οι χρήστες βρίσκονται αρκετά κοντά ο ένας στον άλλον ή βρίσκονται σε παρόμοιες κατευθύνσεις, οι πίνακες καναλιών που αντιστοιχούν σε αυτούς έχουν υψηλή συσχέτιση. Αυτό οδηγεί τους σταθμούς βάσης να αντιλαμβάνονται παρόμοιο χωρικό αποτύπωμα για διαφορετικούς χρήστες, φαινόμενο που οδηγεί σε αποτυχία σύνδεσης με τον σταθμό βάσης[LL08],[SCO13]. (ii) Επειδή οι χρήστες βρίσκονται σε διαφορετικές αποστάσεις από τον BS κατά την επικοινωνία ανερχόμενης ζεύξης (uplink) προκαλείται το φαινόμενο near-far problem. Το φαινόμενο αυτό προκαλεί αστάθεια στον πίνακα καναλιών που παρατηρεί ο BS. Αυτό σημαίνει ότι θα πρέπει να περιοριστεί το εύρος της ισχύος που λαμβάνεται από τους χρήστες. Για το λόγο αυτό, λοιπόν, έχει προταθεί η ομαδοποίηση των κινητών χρηστών σε κλάσεις ισχύος, όπου οι χρήστες κάθε κλάσης δύνανται να χρησιμοποιούν το ίδιο σύνολο καναλιών [Tan95b].

Ένα υποσχόμενο σχήμα SDMA είναι η τεχνική Πολλαπλής Εισόδου - Πολλαπλής Εξόδου (Multiple-input Multiple-output,MIMO) [MK10], [SH10], [Van+00]. Πομποί και δέκτες διαθέτουν πολλαπλές κεραίες. Ο πομπός διαιρεί τις ακολουθίες συμβόλων πληροφορίας σε ανεξάρτητες ροές δεδομένων και μεταδίδει την καθεμία από μία διαφορετική κεραία. Ο κάθε δέκτης στη συνέχεια λαμβάνει έναν ξεχωριστό γραμμικό συνδυασμό των προς μετάδοση διαμορφωμένων ροών (streams). Έχει αποδειχθεί ότι στο σχήμα αυτό η χωρητικότητα αυξάνεται γραμμικά με την αύξηση του πλήθους των κεραιών σε πομπό και δέκτη [LT02]. Το κύριο πλεονέκτηματα του MIMO είναι ότι επιτυγχάνει μεγάλα κέρδη διαδρομών ακόμα και σε περιβάλλοντα πολλών εμποδίων όπου τα φαινόμενα πολυδιαδρομικής διάδοσης είναι έντονα. Επίσης, οι τερματικοί δέκτες είναι επαρκές να διαθέτουν απλές κατασκευές κεραιών, ενώ δεν

Σχήμα 2.3: Απεικόνιση λειτουργίας της τεχνικής NOMA για 2 χρήστες σε downlink μετάδοση [Mao+22].



απαιτείται κάποιος σύνθετος μηχανισμός κατανομής πόρων [Lar+14].

Τα κύρια προβλήματα στον σχεδιασμό συστημάτων SDMA είναι: (i) Η εσφαλμένη ενημέρωση του BS για την κατάσταση των καναλιών (Channel State Information,CSI) περιορίζει τον μέγιστο αριθμό χρηστών που μπορούν να εξυπηρετηθούν στο δίκτυο. (ii) Ο κακός συγχρονισμός BS και τερματικών. Για αυτό τα τελευταία χρόνια τεχνικές όπως η από κοινού εκτίμηση καναλιού (joint channel estimation), αλλά και η ανάπτυξη στατιστικών προκωδικοποιητών, έχουν τραβήξει μεγάλη προσοχή, ώστε να αυξήσουν την ακρίβεια στη μέτρηση του CSI [RHV13], [JH07].

2.2.3 Μη Ορθογωνική Πολλαπλή Πρόσβαση (NOMA)

Υπάρχουν δύο κύριες κατηγορίες τεχνικών NOMA. Στην πρώτη κατηγορία ανήκουν οι τεχνικές που εξετάζουν τη μη ορθογωνική πολλαπλή πρόσβαση στο πεδίο της ισχύος (power-domain NOMA) [Sai+13], ενώ η δεύτερη στο πεδίο του κώδικα code-domain NOMA [NB13]. Η βασική ιδέα των power-domain NOMA σχημάτων είναι η υπέρθεση, στον πομπό, των ιδιωτικών μηνυμάτων των χρηστών στο πεδίο της ισχύος και η ανάκτηση τους στους δέκτες, μέσω διαδοχικής ακύρωσης παρεμβολών (Successive Interference Cancellation,SIC) [Dai+15]. Η τεχνική NOMA εκμεταλλεύεται τις διακυμάνσεις στο κέρδος καναλιού του κάθε χρήστη και αναζητά τις κατάλληλες στάθμες ισχύος για να μεταδώσει ταυτοχρόνως τα μηνύματα, υπό τον περιορισμό να ικανοποιούνται οι συνθήκες της διαδοχικής ακύρωσης παρεμβολών στον κάθε δέκτη [ATH16]. Σε συστήματα πολλαπλών κεραιών, οι χρήστες χωρίζονται σε ομάδες. Κάθε χρήστης αντιμετωπίζει την παρεμβολή από χρήστες ίδιας ομάδας σαν θόρυβο και ανακτά το μήνυμα του με SIC. Μία από τις βασικές αρχές για την ορθή λειτουργία της τεχνικής NOMA είναι να μπορούν οι χρήστες να αποκωδικοποιήσουν όλα τα μηνύματα που προορίζονται σε χρήστες ίδια ομάδας και μικρότερου κέρδους καναλιού. Η παρεμβολή από χρήστες άλλων ομάδων αντιμετωπίζεται κάνοντας χρήση SDMA τεχνικών.

Το πλεονέκτημα της τεχνικής NOMA είναι ότι αυξάνει τη φασματική απόδοση σε συστήματα υψηλής κινητικότητας, σε περιπτώσεις που τα κανάλια των χρηστών παρουσιάζουν διακυμάνσεις. Προκειμένου, όμως, να επιτευχθεί ορθή διαδοχική ακύρωση παρεμβολών, α-

παιτούνται πολύπλοκοι δέκτες. Ειδικότερα, όσο αυξάνεται το πλήθος χρηστών που ανήκουν στην ίδια ομάδα αυξάνονται και τα στρώματα αποκωδικοποίησης που πρέπει να διαθέτουν οι δέκτες. Αν σε ένα σύστημα υπάρχουν K χρήστες, τότε ο χρήστης με το μεγαλύτερο κέρδος καναλιού χρειάζεται $K-1$ στρώματα διαδοχικής ακύρωσης παρεμβολών ώστε να αποκωδικοποιήσει τα μηνύματα από όλους τους χρήστες που έχουν δρομολογηθεί στην ίδια ομάδα και έχουν μικρότερο κέρδος και στη συνέχεια να ανακτήσει το μήνυμα που προορίζεται για εκείνον.

Τα Σχήματα 2.1, 2.2, 2.3 παρέχουν μια σχεδιαστική παρουσίαση του τρόπου λειτουργίας των τεχνικών OMA, SDMA και NOMA για δύο χρήστες σε downlink μετάδοση αντίστοιχα και κατ΄ επέκταση μπορούν να οδηγήσουν σε σύγκριση των τεχνικών αυτών μεταξύ τους. Συγκεκριμένα, σύμφωνα με την τεχνική OMA, ο ένας χρήστης καταλαμβάνει ολόκληρο τον ραδιοπόρο, ενώ ο δεύτερος χρήστης πρέπει να περιμένει να ελευθερωθεί ο ραδιοπόρος αυτός, ώστε να αποκτήσει πρόσβαση στο δίκτυο. Κατά την τεχνική SDMA, τα δύο μηνύματα κωδικοποιούνται σε δύο ανεξάρτητα streams, τα οποία στη συνέχεια υφίστανται γραμμική προκωδικοποίηση στον πομπό. Ο δέκτης κάθε χρήστη αποκωδικοποιεί απευθείας το stream που προορίζεται για εκείνον αντιμετωπίζοντας το άλλο stream σαν θόρυβο. Στην περίπτωση της τεχνικής NOMA, τα δύο μηνύματα κωδικοποιούνται σε ανεξάρτητα streams τα οποία μεταδίδονται προς όλους τους χρήστες. Ο χρήστης με μεγαλύτερο κέρδος καναλιού αποκωδικοποιεί και τα δύο μηνύματα.

2.2.4 Πολλαπλή Πρόσβαση Διαίρεσης Ρυθμού (RSMA)

Λόγω της φύσης των ασύρματων καναλιών, η διαχείριση της παρεμβολής μεταξύ πολλαπλών χρηστών είναι ζωτικής σημασίας στη σχεδίαση ασύρματων δικτύων. Τα προηγούμενα σχήματα πολλαπλής πρόσβασης υιοθετούν διαφορετική προσέγγιση στον τρόπο που διαχειρίζονται την παρεμβολή. Στο σχήμα OMA, η παρεμβολή αποφεύγεται με ανάθεση στους χρήστες ορθογωνικών μεταξύ τους ραδιοπόρων ενώ στο σχήμα SDMA η παρεμβολή αντιμετωπίζεται ως θόρυβος. Η εξέλιξη των σχημάτων NOMA επέτρεψε στους χρήστες να αποκωδικοποιούν την παρεμβολή. Τα σχήματα πολλαπλής πρόσβασης νέας γενιάς οραματίζονται μία προσαρμοστική προσέγγιση κατά την οποία κάποιο μέρος της παρεμβολής θα αποκωδικοποιείται στο δέκτη ενώ το υπόλοιπο θα αντιμετωπίζεται σαν θόρυβος ανάλογα με τις συνθήκες που επικρατούν στο ασύρματο μέσο.

Η τεχνική RSMA είναι μία πολλά υποσχόμενη τεχνική πολλαπλής πρόσβασης ο σχεδιασμός της οποίας στοχεύει στην αξιοποίηση των πλεονεκτημάτων των SDMA και NOMA σχημάτων.

Στην παρούσα εργασία εστιάζουμε αποκλειστικά στην downlink μετάδοση. Ανά διαστήματα έχουν προταθεί αρκετά RSMA σχήματα. Στις εργασίες [CC15] - [MPC21] έχουν προταθεί σχήματα ενός επιπέδου, ενώ στην εργασία [JCS20] έχει προταθεί ένα ιεραρχικό RS σχήμα 2 επιπέδων. Τέλος στις εργασίες [Li+20] και [Ahm+20] προτείνονται γενικευμένα πλαίσια RS σχημάτων, ενώ στις δημοσιεύσεις [Ahm+19], [MC20a], [MC20b] προτείνονται τεχνικές για κωδικοποίηση των κοινών μηνυμάτων.

Αρχιτεκτονική Πομπού

Σε όλα τα downlink RSMA σχήματα ο πομπός ενός σταθμού βάσης μπορεί να αναπαρασταθεί με τη διάταξη του Σχήματος 2.4. Ο πομπός διαθέτει $M \geq 1$ κεραίες και εξυπηρετεί K χρήστες. Ο σχεδιασμός είναι τέτοιος ώστε να επιτυγχάνεται καλύτερος έλεγχος της διακαναλικής παρεμβολής.

Σε αντίθεση με τα συμβατικά σχήματα πολλαπλής πρόσβασης οι πομποί είναι εφοδιασμένοι με ένα στοιχείο που ονομάζεται Message Splitter (Διαχωριστής Μηνύματος). Η λειτουργία του στοιχείου αυτού είναι να διαχωρίζει τα μηνύματα W_k των χρηστών σε L υπο-μηνύματα $\{W_k^1, W_k^2, \ldots, W_k^L\}$ όπου k είναι ο δείκτης χρήστη. Το πλήθος των υπο-μηνυμάτων εξαρτάται από το προτεινόμενο σχήμα RSMA. Για παράδειγμα το μήνυμα κάθε χρήστη σε σχήματα ενός επιπέδου διαχωρίζεται σε 2 υπομηνύματα, σε σχήματα 2 επιπέδων σε 3 υπομηνύματα ενώ στα γενικευμένα σχήματα RS, $L = 2^{K-1}$.

Στη συνέχεια τα υπο-μηνύματα συνδυάζονται μεταξύ τους. Για τη διαδικασία αυτή είναι υπεύθυνο το στοιχείο Message Combiner (Συνδυαστής Μηνυμάτων) που απεικονίζεται στο Σχήμα 2.4. Η έξοδος του Message Combiner είναι N μηνύματα $W = \{W_1^*, W_2^*, \ldots, W_N^*\}$ όπου $N = K + 1$ για σχήματα με 1 επίπεδο, $N = K + 2$ για σχήματα 2 επιπέδων και $N = 2^K - 1$ για γενικευμένα πλαίσια RS. Όταν κάποιο από τα N μηνύματα προορίζεται για όλους τους χρήστες του συστήματος ονομάζεται κοινό μήνυμα και συνήθως αποτελείται από υπομηνύματα διαφορετικών χρηστών. Τα μηνύματα που προορίζονται για έναν μόνο χρήστη καλούνται ιδιωτικά και αποτελούνται από υπομηνύματα του ίδιου του χρήστη.

Κάθε κοινό μήνυμα κωδικοποιείται σε ένα κοινό stream με χρήση ενός βιβλίου κωδικών (codebook) που είναι γνωστό σε όλους τους χρήστες. Στον αντίποδα κάθε ιδιωτικό μήνυμα κωδικοποιείται σε ανεξάρτητο stream με χρήση ενός μόνο κωδικού που είναι γνωστός μόνο στον χρήστη για τον οποίο προορίζεται το μήνυμα. Στη συνέχεια N προκωδικοποιητές (precoders) αντιστοιχούν τις N ροές δεδομένων $s_1, s_2, \ldots, s_N$ σε κατάλληλα διανύσματα διαμόρφωσης δέσμης (beamforming vectors) P_1,P_2,...,P_N , όπου $P_n \in \mathbb{C}^{M\times 1}$, ώστε να μεταδοθούν από τις M κεραίες εκπομπής που διαθέτει ο πομπός. Το τελικό σήμα εκπομπής είναι η υπέρθεση όλων των κοινών και ιδιωτικών streams και εκφράζεται μαθηματικά ως:

$$x = \sum_{n=1}^{N} P_n s_n. \tag{2.1}$$

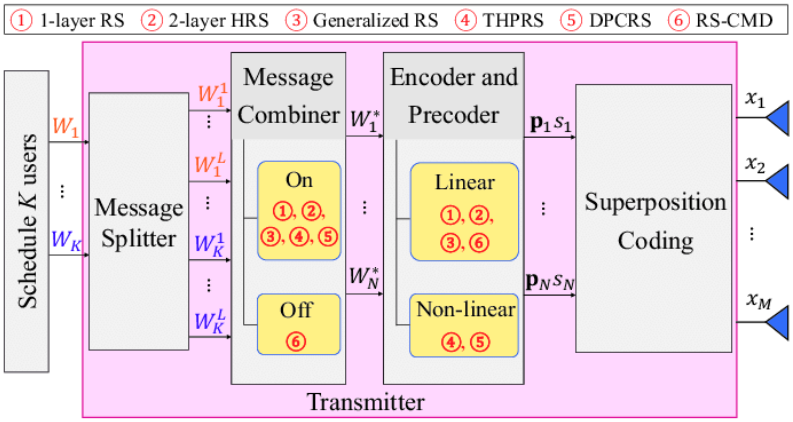
Αρχιτεκτονική Δέκτη

Η αρχιτεκτονική ενός δέκτη σε RSMA σχήμα απεικονίζεται στο Σχήμα 2.5. Ο δέκτης k λαμβάνει το σήμα y_k:

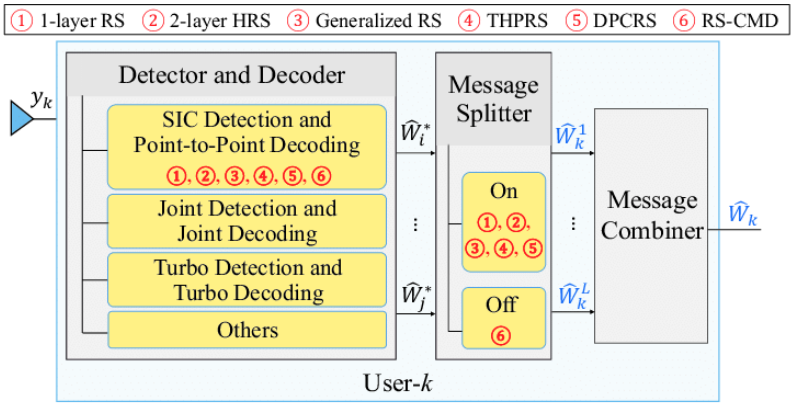
$$y_k = g_k^H x + n_k, \tag{2.2}$$

όπου $g_k \in \mathbb{C}^{M\times 1}$ είναι το διάνυσμα καναλιού του χρήστη k και n_k είναι ο Αθροιστικός Λευκός Θόρυβος Gauss (AWGN). Το σήμα y_k διέρχεται από αποκωδικοποιητή ώστε ο δέκτης να ανακτήσει μια εκτίμηση $\hat{W}_k$ του μηνήματος W_k. Σύμφωνα με την στρατηγική κωδικοποίησης

Σχήμα 2.4: Downlink RSMA Πομπός [Mao+22].



Σχήμα 2.5: Downlink RSMA Δέκτης [Mao+22].



που εφαρμόζει το χρησιμοποιούμενο από το δίκτυο σχήμα RSMA, κάθε χρήστης αποκωδικοποιεί ένα υποσύνολο των εκπεμπόμενων μηνυμάτων $\hat{W}_i^*, \ldots, \hat{W}_j^*$. Στη συνέχεια το στοιχείο Message Splitter ανακτά τα υπο-μηνύματα $\hat{W}_k^1, \hat{W}_k^2, \ldots, \hat{W}_k^L$ που ανήκουν στον χρήστη k και στο τελικό στάδιο ένας Combiner ανακτά το αρχικό μήνυμα $\hat{W}_k$.

2.2.5 Πλεονεκτήματα RSMA

Οι τεχνικές RSMA παρουσιάζουν κάποια πλεονεκτήματα σε σχέση με τα σχήματα SDMA, NOMA και OMA που χρησιμοποιούνται στα σύγχρονα ασύρματα δίκτυα. Τα πλεονεκτήματα συνοψίζονται παρακάτω:

- **Ευελιξία**: Το σχήμα RSMA είναι κατάλληλο για μεταβλητό φόρτο δικτύου και διαφορετικές απαιτήσεις χρηστών. Σε αντίθεση με την τεχνική SDMA που αντιμετωπίζει την παρεμβολή με μηχανισμούς στην πλευρά του πομπού και την NOMA που διαθέτει μηχανισμούς αποκωδικοποίησης της παρεμβολής στους δέκτες, το RSMA κατασκευάζει κοινά streams που επιτρέπουν τη διαχείριση της παρεμβολής στον πομπό αλλά και στον

δέκτη.

- **Ανθεκτικότητα**: Η RSMA είναι ανθεκτική σε σφάλματα που συμβαίνουν κατά τη διάρκεια της μέτρησης του CSI. Τα σφάλματα οφείλονται σε αρκετούς λόγους, όπως το pilot contamination [Mis+22b] που εμφανίζεται στα massive MIMO συστήματα, δηλαδή συστήματα με τεράστιες συστοιχίες κεραιών στους πομπούς. Η κύρια αιτία όμως που προκαλεί εσφαλμένη ενημέρωση του CSI είναι η κινητικότητα των χρηστών στο δίκτυο [DMC21].

- **Υψηλή Φασματική Απόδοση**

- **Υψηλή Ενεργειακή Απόδοση**

- **Επέκταση Κάλυψης**: Στην εργασία [Mao+20], έχει αποδειχθεί ότι ο ρυθμός μετάδοσης δεδομένων των χρηστών μπορεί να εξισορροπηθεί μέσα από ένα min-max παιχνίδι ακόμα και όταν τα κέρδη καναλιών των χρηστών παρουσιάζουν μεγάλες διαφορές.

- **Χαμηλή Καθυστέρηση Απόκρισης**

Παρόλο που το RSMA παρουσιάζει τα πλεονεκτήματα που αναφέρθηκαν παραπάνω, αναπόφευκτα προκαλούνται ορισμένα προβλήματα. Το κύριο πρόβλημα είναι οι περιορισμοί SIC που καλούνται να διαχειριστούν οι δέκτες ώστε να μπορούν να αποκωδικοποιήσουν ορθά τα κοινά μηνύματα. Συγκεκριμένα, όσο αυξάνονται τα επίπεδα αποκωδικοποίησης απαιτούνται πιο σύνθετοι δέκτες. Ένα άλλο πρόβλημα που προκύπτει είναι ότι το πλήθος των streams που καλείται να κωδικοποιήσει ο πομπός είναι μεγαλύτερο από το πλήθος των χρηστών. Επομένως, δημιουργείται η ανάγκη για όλο και πιο σύνθετους κωδικοποιητές ώστε οι ροές δεδομένων να κωδικοποιηθούν σε κατάλληλα διανύσματα μορφοποίησης δέσμης (beamforming vectors). Επίσης, για είναι ο δέκτης ικανός να συνθέσει το αρχικό σήμα από τα υπο-μηνύματα που έχει ανακτήσει χρειάζεται να γνωρίζει ποια μηνύματα να αποκωδικοποιήσει αλλά και πως να τα συνδυάσει στον Message Combiner. Αυτό οδηγεί σε απαιτητικό συγχρονισμό μεταξύ πομπού και δέκτη.

2.2.6 Προκλήσεις και Μελλοντικές Έρευνες

Από την παραπάνω ανάλυση, αποδεικνύεται πως τα ασύρματα δίκτυα που υποστηρίζονται από σχήματα RSMA παρουσιάζουν σημαντικά οφέλη με αποτέλεσμα να έχει δημιουργηθεί σοβαρή ερευνητική δραστηριότητα γύρω από αναδυόμενες εφαρμογές στα δίκτυα νέας γενιάς. Ωστόσο, η έρευνα των RSMA τεχνικών βρίσκεται ακόμα σε πρώιμο στάδιο καθώς υπάρχουν πολλά ανοικτά προβλήματα και ερευνητικές κατευθύνσεις που θα πρέπει να μελετηθούν από την επιστημονική κοινότητα. Μερικές από αυτές είναι:

- Η μελέτη του RSMA σε συστήματα massive MIMO και ειδικότερα σε cell-free δίκτυα [Mis+22a].

- Επιστράτευση τεχνικών συμπιεσμένης ανίχνευσης και βαθιάς μάθησης σε mmWave δίκτυα (δίκτυα που εκπέμπουν σε συχνότητες άνω των 30GHz) με πολλαπλή πρόσβαση RSMA, με σκοπό να συμπιεστεί το CSI σε πίνακες χαμηλότερων διαστάσεων [DC17], [ALH21].

- Η επίδραση του RSMA σε συνθήκες multicast εκπομπής [Ter+18a], [Ter+18b], [YC21], [YYC20],[Che+21].

- Εξερεύνηση του RSMA σε δίκτυα πολλαπλών κυψελών με αρχιτεκτονική νέφους (C-RAN) ή αρχιτεκτονική ομίχλης (F-RAN) [Has+21b], [Rei+21], [Has+21a].

Υπάρχουν επίσης πολλές αναδυόμενες τεχνολογίες στον τομέα των τηλεπικοινωνιών όπου η επιστημονική κοινότητα έχει στρέψει το ενδιαφέρον της τα τελευταία χρόνια. Υπάρχει, επομένως, μεγάλο ενδιαφέρον να μελετηθούν τα οφέλη και οι επιπτώσεις ενός σχήματος διαίρεσης ρυθμού στις εφαρμογές αυτές. Κάποιες από αυτές είναι:

- Η τεχνολογία Simultaneous Wireless Information and Power Transfer (SWIPT) που στοχεύει στην ταυτόχρονη μεταφορά πληροφορίας και ενέργειας στους δέκτες [AMK21], [Su+19].

- Η χρήση μη επανδρωμένων εναέρων οχημάτων (Unmanned Aerial Vehicle-UAV) ώστε να αυξηθεί η περιοχή κάλυψης [Jaa+20a], [Jaa+20b].

- Ασφάλεια φυσικού στρώματος (Physical Layer-PHY) [DC21].

- Δορυφορικές Επικοινωνίες [YDC21], [Lin+21].

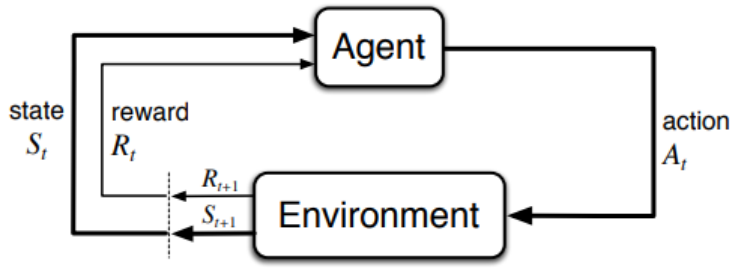
2.3 Βασικές Έννοιες Ενισχυτικής Μάθησης

Η Ενισχυτική Μάθηση (Reinforcement Learning) είναι μια κλάση μεθόδων επίλυσης προβλημάτων διαδοχικής λήψης αποφάσεων με στόχο τη μεγιστοποίηση ενός σωρευτικού αριθμητικού σήματος ανταμοιβής (reward signal).

Η βασική ιδέα είναι ότι χρησιμοποιείται ένας πράκτορας μάθησης που αλληλεπιδρά με το περιβάλλον του προβλήματος. Ο πράκτορας αντιλαμβάνεται κάποια χαρακτηριστικά του περιβάλλοντος και επιλέγει ενέργειες οι οποίες το επηρεάζουν. Επίσης, ο πράκτορας θα πρέπει να έχει τουλάχιστον έναν στόχο επίτευξης ως προς κάποια από τα χαρακτηριστικά που παρατηρεί.

Μία από τις βασικές προκλήσεις της ενισχυτικής μάθησης είναι το trade-off (εξισορρόπηση) μεταξύ εκμετάλλευσης και εξερεύνησης της γνώσης που αποκτά ο πράκτορας. Προκειμένου να μεγιστοποίησει το σήμα ανταμοιβής, ο πράκτορας προτιμά ενέργειες που ήδη γνωρίζει ότι οδηγούν σε υψηλή ανταμοιβή. Προκειμένου, όμως, να ανακαλύψει τις ενέργειες αυτές, θα πρέπει να δοκιμάσει καινούργιες οι οποίες πολλές φορές μπορεί να επιστρέφουν χαμηλότερη ανταμοιβή. Επομένως, θα πρέπει στην αρχή να δοκιμάζει καινούργιες ενέργειες και σταδιακά να επιλέγει αυτές που είναι καλύτερες. Ειδικότερα σε περιβάλλοντα που μεταβάλλονται στοχαστικά, κάθε ενέργεια πρέπει να επιλεχθεί πολλές φορές ώστε ο πράκτορας να μάθει αν η

Σχήμα 2.6: Απεικόνιση Διεπαφής Πράκτορα-Περιβάλλοντος [SB18]



ενέργεια αυτή είναι πραγματικά ωφέλιμη. Ανάλογα, λοιπόν, με την πολυπλοκότητα του περιβάλλοντος, το πρόβλημα εκμετάλλευσης-εξερεύνησης (exploration-exploitation) παρουσιάζει αυξημένη δυσκολία.

2.3.1 Στοιχεία Ενισχυτικής Μάθησης

Στην ενότητα αυτή περιγράφουμε το πλαίσιο στο οποίο ένα πρόβλημα μπορεί να μοντελοποιηθεί ως πρόβλημα ενισχυτικής μάθησης. Εισάγουμε τα βασικά στοιχεία ενός τέτοιου πλαισίου και διατυπώνουμε τους αναγκαίους ορισμούς που θα μας χρειαστούν στη συνέχεια.

Διεπαφή Πράκτορα-Περιβάλλοντος

Υποθέτουμε ότι ο πράκτορας και το περιβάλλον αλληλεπιδρούν ανά διακριτά χρονικά βήματα $t = 0, 1, 2, \ldots$. Στο χρονικό βήμα t, ο πράκτορας παρατηρεί κάποια χαρακτηριστικά του περιβάλλοντος και σχηματίζει μία περιγραφή αυτού που ονομάζεται κατάσταση (state). Η κατάσταση συμβολίζεται με $S_t \in S$, όπου S είναι το σύνολο όλων των δυνατών καταστάσεων που μπορεί να παρατηρήσει ο πράκτορας. Με βάση την κατάσταση στην οποία βρίσκεται, ο πράκτορας επιλέγει και εκτελεί μία ενέργεια $A_t \in A_{S_t}$, όπου A_{S_t} είναι το σύνολο των δυνατών ενεργειών που μπορεί να επιλέξει ο πράκτορας όταν παρατηρεί την κατάσταση S_t. Στο επόμενο χρονικό βήμα, ως συνέπεια της ενέργειας που επέλεξε, ο πράκτορας μεταβαίνει σε μία νέα κατάσταση S_{t+1} και λαμβάνει ένα βαθμωτό σήμα ανταμοιβής $R_{t+1} \in R$. Η διαδικασία αυτή περιγράφεται στο Σχήμα 2.6.

Η επιλογή της ενέργειας από τον πράκτορα γίνεται με βάση την πολιτική π που υιοθετεί. Η πολιτική π είναι μια απεικόνιση τους ζεύγους (A_t, S_t) σε μία πιθανότητα που συμβολίζεται με $\pi(A_t/S_t)$ και περιγράφει την δεσμευμένη πιθανότητα να επιλεγεί η ενέργεια A_t δεδομένου ότι η κατάσταση του πράκτορα είναι S_t. Αυτό που στην ουσία περιγράφει μία μέθοδος ενισχυτικής μάθησης είναι ο τρόπος με τον οποίο ο πράκτορας μεταβάλλει την πολιτική π καθώς αποκτά εμπειρία από το περιβάλλον.

Ανταμοιβή

Στην ενισχυτική μάθηση, ο στόχος του πράκτορα μοντελοποιείται με κατάλληλη επιλογή του σήματος ανταμοιβής. Όταν μια πολιτική διαμορφώνει μια ακολουθία ενεργειών που οδηγεί

τον πράκτορα πιο κοντά στον στόχο του, η ακολουθία των σημάτων ανταμοιβής πρέπει να είναι «καλύτερη» σε σχέση με μια πολιτική που τον απομακρύνει από αυτόν. Δηλαδή, το σήμα ανταμοιβής είναι ένας τρόπος να ενημερώνεται ο πράκτορας για το πόσο ωφέλιμη είναι η πολιτική που ακολουθεί. Ισοδύναμα, ο στόχος του πράκτορα είναι να βρει την πολιτική εκείνη που μεγιστοποίει ένα καλώς ορισμένο μέγεθος G_t που εξαρτάται από τα σήματα ανταμοιβής που λαμβάνει σε βάθος χρόνου.

Έστω η ακολουθία ανταμοιβών $R_{t+1}, R_{t+2}, R_{t+3}, \ldots$. Στην πιο απλή περίπτωση, το μέγεθος που επιθυμεί ο πράκτορας να μεγιστοποιήσει ορίζεται ως:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots . \tag{2.3}$$

Το άθροισμα αυτό είναι άπειρο για προβλήματα γνωστά ως continuing tasks, στα οποία η διεπαφή πράκτορα-περιβάλλον δεν «σπάει». Στα περισσότερα, όμως, πρακτικά προβλήματα, όταν ο πράκτορας μεταβεί σε μια ειδική κατάσταση που ονομάζεται τερματική (terminal state), η διεπαφή «σπάει» και ο πράκτορας μεταβαίνει σε μια αρχική κατάσταση (τυχαία ή ντετερμινιστικά). Τα προβλήματα αυτά είναι γνωστά ως episodic tasks και λέμε ότι το σύστημα από την αρχική κατάσταση μέχρι να βρεθεί στην τερματική διαγράφει ένα επεισόδιο. Αυτή η ιδιότητα μας επιτρέπει να χειριστούμε πεπερασμένες ακολουθίες $R_{t+1}, R_{t+2}, R_{t+3}, \ldots, R_T$.

Μία επιπλέον έννοια που συναντάται στη βιβλιογραφία είναι ο όρος "discounting". Η πολιτική που ακολουθεί ένας πράκτορας στο βήμα t έχει μεγαλύτερη επίδραση στις ανταμοιβές που θα λάβει από τα κοντινότερα του επόμενα βήματα. Για αυτόν το λόγο, εισάγουμε την παράμετρο γ γνωστή ως discounted rate. Η παράμετρος γ καθορίζει πόσο σημαντικές είναι οι μελλοντικές ανταμοιβές και παίρνει τιμές στο διάστημα $[0, 1]$. Το μέγεθος G_t ξαναορίζεται ως:

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}, \tag{2.4}$$

όπου T είναι το μήκος του επεισοδίου. Όταν $\gamma = 0$, λαμβάνεται υπόψιν μόνο η άμεση ανταμοιβή, ενώ όσο η παράμετρος γ πλησιάζει τη μονάδα οι μελλοντικές ανταμοιβές έχουν μεγαλύτερη βαρύτητα. Το μέγεθος G_t ονομάζεται discounted return.

2.3.2 Μαρκοβιανές Διαδικασίες Απόφασης (MDP)

Ένα πρόβλημα ενισχυτικής μάθησης λέμε ότι έχει την ιδιότητα Markov όταν ικανοποιείται η ακόλουθη σχέση:

$$Pr\{R_{t+1} = r, S_{t+1} = s^{'} | S_0, A_0, R_1, \ldots, S_{t-1}, A_{t-1}, R_t, S_t, A_t\} = Pr\{R_{t+1} = r, S_{t+1} = s^{'} | S_t, A_t\}, \tag{2.5}$$

δηλαδή η δυναμική του προβλήματος είναι τέτοια ώστε η πιθανότητα μετάβασης στην νέα κατάσταση $s^{'}$ με ταυτόχρονη λήψη ανταμοιβής r, εξαρτάται μόνο από την τωρινή κατάσταση S_t και την πιο πρόσφατη ενέργεια του πράκτορα, A_t.

Ένα πρόβλημα ενισχυτικής μάθησης μπορεί να περιγραφεί ως MDP και αποτελείται από τα εξής στοιχεία:

- Ένα σύνολο καταστάσεων S και μια κατανομή αρχικών καταστάσεων $p(s_0)$.

- Τα σύνολα ενεργειών A_{S_t}. Υποθέτουμε για απλούστευση ότι σε κάθε κατάσταση σε οποιοδήποτε βήμα t, το σύνολο των επιτρεπτών ενεργειών είναι το ίδιο και συμβολίζεται με A.

- Τη δυναμική του περιβάλλοντος $T(S_{t+1}|S_t, A_t)$ που αντιστοιχίζει κάθε ζευγάρι (κατάσταση − ενέργεια) του βήματος t σε μια κατανομή καταστάσεων για το βήμα $t+1$.

- Τη συνάρτηση άμεσης ανταμοιβής $R_{t+1} = R(S_t, A_t, S_{t+1})$. Πολλές φορές η ανταμοιβή δεν εξαρτάται από την επόμενη κατάσταση S_{t+1}.

- Τον discount factor γ που καθορίζει τη βαρύτητα των μελλοντικών αμοιβών.

Ο σκοπός της ενισχυτικής μάθησης είναι να βρει μια πολιτκή π^* που να μεγιστοποιεί το μέγεθος G_t για όλες τις καταστάσεις. Συγκεκριμένα:

$$\pi^* = \arg\max_{\pi} \mathbb{E}[G_t|\pi, S_t = s], \quad \forall s \in S. \tag{2.6}$$

2.3.3 Συναρτήσεις Τιμής (Value Functions)

Βασικό στοιχείο κάθε αλγορίθμου ενισχυτικής μάθησης είναι η *συνάρτηση τιμής (value function)* των καταστάσεων ή των ζευγών κατάσταση-ενέργεια. Η *συνάρτηση τιμής* είναι ένα μέτρο του πόσο «καλό» είναι να βρεθεί σε μία κατάσταση s ο πράκτορας. Το «καλό» μετράται σε σχέση με την *αναμενόμενη επιστροφή* (G_t) όταν η τωρινή κατάσταση είναι η s. Για MDP προβλήματα, η *τιμή* της κατάστασης s δεδομένης μιας πολιτικής π ορίζεται ως:

$$\upsilon_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} / S_t = s \right]. \tag{2.7}$$

Σημειώνουμε ότι η *τιμή* μιας τερματικής κατάστασης είναι μηδενική. Η συνάρτηση υ_π ονομάζεται state-value function της πολιτικής π.

Όμοια μπορούμε να ορίσουμε την state-action value function της πολιτικής π. Η state-action value function περιγράφει πόσο «καλή» είναι μία ενέργεια α δεδομένου ότι η κατάσταση του πράκτορα είναι s και ορίζεται ως:

$$Q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} / S_t = s, A_t = a \right]. \tag{2.8}$$

Μία σημαντική ιδιότητα των *συναρτήσεων τιμής* είναι ότι ικανοποιούν αναδρομικές σχέσεις. Μία από τις πιο σημαντικές σχέσεις στην ενισχυτική μάθηση είναι η εξίσωση του **Bellman** [Men+20] η οποία περιγράφει την σχέση της *τιμής* μιας κατάστασης s με τις *τιμές* όλων των καταστάσεων $s^{'}$ που είναι πιθανόν να διαδεχθούν την s. Συγκεκριμένα, για την συνάρτηση $Q_\pi(s, a)$ ισχύει:

$$Q_\pi(S_t, A_t) = \mathbb{E}_{S_{t+1}}[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi(S_{t+1}))]. \tag{2.9}$$

Οι *συναρτήσεις τιμής* ορίζουν σχέση μερικής διάταξης στο σύνολο των πολιτικών. Μια πολιτική π ορίζεται ότι είναι καλύτερη ή ίση με την πολιτική $\pi^{'}$ και συμβολίζεται με $\pi \geq \pi^{'}$ αν και μόνο αν $\upsilon_{\pi}(s) \geq \upsilon_{\pi^{'}}(s), \quad \forall s \in S$. Το σύνολο των πολιτικών π_* για το οποίο ισχύει $\upsilon_{\pi_*}(s) \geq \upsilon_{\pi(s)}, \quad \forall s \in S$ και $\forall \pi$ ονομάζεται σύνολο βέλτιστων πολιτικών. Όλες οι βέλτιστες πολιτικές έχουν τις ίδιες state value function και state-action value function και ορίζονται ως:

$$\upsilon_*(s) = \max_{\pi} \upsilon_{\pi}(s), \quad \forall s \in S, \tag{2.10}$$

και

$$Q^*(s,a) = \max_{\pi} Q_{\pi}(s,a), \quad \forall s \in S \quad \text{και} \quad \forall a \in A, \tag{2.11}$$

αντίστοιχα.

Η εξίσωση 2.9 είναι η βάση των μεθόδων Q-learning [WD92] και SARSA [RN94] :

$$Q_{\pi}(S^t, A^t) \leftarrow Q_{\pi}(S^t, A^t) + \alpha\delta, \tag{2.12}$$

όπου α είναι ο ρυθμός μάθησης και $\delta = Y - Q_{\pi}(S^t, A^t)$ είναι το σφάλμα προσωρινής διαφοράς (Temporal Difference-TD). Το Y είναι μία τιμή στόχος (target) η οποία εξαρτάται από τον αλγόριθμο που χρησιμοποιείται. Η τιμή Υ υπολογίζεται με βάση την πολιτική στόχος (target policy). Η target policy είναι η πολιτική που προσπαθεί να μάθει ο πράκτορας, δηλαδή μαθαίνει τη *συνάρτηση τιμής* που αντιστοιχεί στην πολιτική αυτή. Η πολιτική συμπεριφοράς (behavior policy) είναι η πολιτική που χρησιμοποιείται από τον πράκτορα για επιλογή ενέργειας, δηλαδή μέσω αυτής ο πράκτορας αλληλεπιδρά με το περιβάλλον.

Ο SARSA είναι ένας on-policy αλγόριθμος, δηλαδή βελτιώνει την τωρινή του εκτίμηση για την συνάρτηση Q^* παρατηρώντας τις μεταβάσεις που προκύπτουν από την target policy. Συνεπώς η target policy ταυτίζεται με την behavior policy και $Y = R_t + Q_{\pi}(S^t, A^t)$.

Ο Q-learning σε αντίθεση με τον SARSA είναι ένας off-policy αλγόριθμος, δηλαδή βελτιώνει την τωρινή του εκτίμηση για την συνάρτηση Q_{π} παρατηρώντας μεταβάσεις που προκύπτουν από την behavior policy η οποία δεν ταυτίζεται κατά ανάγκη με την target policy. Στο Q-learning ισχύει $Y = R_t + \max_a Q_{\pi}(S^t, \alpha)$, δηλαδή προσπαθεί να προσεγγίσει απευθείας την Q^* παρατηρώντας τη μετάβαση που οφείλεται στη βέλτιστη ενέργεια για την κατάσταση S_t.

Για να βρεθεί το Q^* από μία τυχαία Q_{π} χρησιμοποιείται η γενικευμένη μέθοδος επανάληψης πολιτικής (policy iteration). Σε πρώτο βήμα, επιστρατεύονται μέθοδοι για την εκτίμηση της συνάρτησης Q_{π} (policy evaluation) και σε δεύτερη φάση, η Q_{π} ενημερώνεται μέσω ελαχιστοποίησης των σφαλμάτων προσωρινής διαφοράς που προκύπτουν από την παρατήρηση των μεταβάσεων και των ανταμοιβών (policy update).

Οι πιο απλές μέθοδοι ενισχυτικής μάθησης αναπαριστούν τις συναρτήσεις τιμής με πίνακες (tabular methods). Υπάρχει μία καταχώρηση στον πίνακα για κάθε κατάσταση ή για κάθε ζεύγος κατάσταση-ενέργεια. Οι μέθοδοι αυτές, όμως, περιορίζονται σε προβλήματα με μικρό πλήθος καταστάσεων και ενεργειών. Πέρα από την περιορισμένη μνήμη, ένα βασικό πρόβλημα που προκύπτει στις tabular μεθόδους είναι η αδυναμία να γενικευτεί η γνώση που έχει

αποκτηθεί σε ένα μικρό υποσύνολο του χώρου καταστάσεων σε όλόκληρο τον χώρο. Ειδικότερα, όταν ο χώρος καταστάσεων είναι συνεχής χρειαζόμαστε μεθόδους που να γενικεύουν καλύτερα τη γνώση.

Η καινοτομία στις μεθόδους Προσέγγισης Συνάρτησης Τιμής (Value Function Approximation) είναι ότι αναπαριστούν τις συναρτήσεις τιμής ως παραμετροποιημένες συναρτήσεις κάποιας μορφής με διάνυσμα παραμέτρων $\mathbf{w} \epsilon R^n$. Η βασική ιδέα είναι ότι η γνώση αποθηκεύεται στο διάνυσμα $\mathbf{w}$ και η ενημέρωση του Q_π γίνεται μέσω της ενημέρωσης του $\mathbf{w}$. Μία από τις πιο γνωστές μεθόδους Προσέγγισης Συνάρτησης Τιμής είναι η χρήση Νευρωνικών Δικτύων (Neural Networks) για την αναπαράσταση της συνάρτησης τιμής, τα οποία μας επιτρέπουν να αναπαραστήσουμε χώρους μεγάλων διαστάσεων.

2.3.4 Αναζήτηση Πολιτικής (Policy Search)

Οι μέθοδοι αναζήτησης πολιτικής δεν διατηρούν κάποιο value function μοντέλο, αλλά αναζητούν απευθείας τη βέλτιστη πολιτική π^*. Τυπικά επιλέγεται μία παραμετροποιημένη πολιτική π_w, της οποίας οι παράμετροι ενημερώνονται ώστε να μεγιστοποιηθεί το $\mathbb{E}[G_t|\pi]$ χρησιμοποιώντας κατάλληλη τεχνική ενημέρωσης. Έχουν προταθεί πολλές τεχνικές ενημέρωσης της παραμέτρου w. Κάποιες από αυτές βασίζονται στη μέθοδο της κλίσης (gradient-based) [Wie+10], [Wil92], ενώ άλλες όχι (gradient-free) [Cuc+11]. Σε περίπτωση που χρησιμοποιούνται νευρωνικά δίκτυα για την παραμετροποίηση της πολιτκής, προτιμώνται οι gradient-based μέθοδοι, καθώς έχει αποδειχθεί ότι μπορούν να χειριστούν καλύτερα πολιτικές πολλών παραμέτρων [Kou+13].

Γενικά, όταν χρησιμοποιούνται παραμετροποιημένες συναρτήσεις τιμής είτε για την αναπαράσταση της state-action value function είτε για την απευθείας αναπαράσταση της πολιτικής, ορίζεται μία συνάρτηση σφάλματος $L(\mathbf{w})$ πάνω σε ένα σύνολο δειγμάτων εκπαίδευσης (training samples). Σε κάθε βήμα εκπαίδευσης, οδηγούμαστε σε μια καλύτερη εκτίμηση της παραμέτρου $\mathbf{w}$. Από τις πιο ευρέως χρησιμοποιούμενες τεχνικές ενημέρωσης των παραμέτρων είναι η τεχνική ανοδικής ή καθοδικής κλίσης (gradient ascent/descent), η οποία βασίζεται στον υπολογισμό των μερικών παραγώγων της συνάρτησης σφάλματος ως προς τις συνιστώσες του $\mathbf{w}$. Η ενημέρωση στην περίπτωση της καθοδικής κλίσης γίνεται ως εξής:

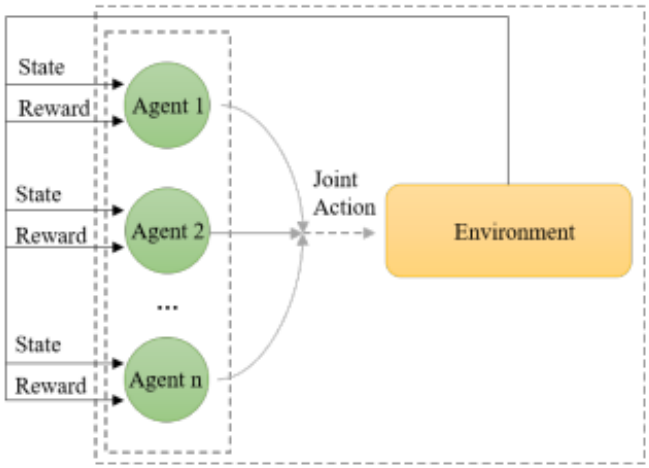
$$\mathbf{w} = \mathbf{w} - \eta \nabla_{\mathbf{w}} L(\mathbf{w}), \tag{2.13}$$

όπου η είναι ο ρυθμός μάθησης. Η ιδέα είναι ότι η παράμετρος $\mathbf{w}$ σε κάθε βήμα ενημέρωσης μετατοπίζεται αντίθετα της κατεύθυνσης που ορίζεται από τον όρο $\nabla L(\mathbf{w})$. Ο λόγος είναι ότι, γεωμετρικά, ο όρος $\nabla L(\mathbf{w})$ περιγράφει την κατεύθυνση μέγιστου ρυθμού αύξησης του $L(\mathbf{w})$ ως προς το $\mathbf{w}$. Στην περίπτωση της ανοδικής κλισης ισχύει:

$$\mathbf{w} = \mathbf{w} + \eta \nabla_{\mathbf{w}} L(\mathbf{w}), \tag{2.14}$$

δηλαδή το $\mathbf{w}$ μετατοπίζεται προς την κατεύθυνση μέγιστου ρυθμού αύξησης του $L(\mathbf{w})$. Η μέθοδος της ανοδικής κλίσης χρησιμοποιείται για προβλήματα εύρεσης μεγίστων.

Σχήμα 2.7: Πολλαπλοί πράκτορες αλληλεπιδρούν με το ίδιο περιβάλλον [NNN19]



2.3.5 Συστήματα Πολλαπλών Πρακτόρων (Multi-Agent Systems-MASs)

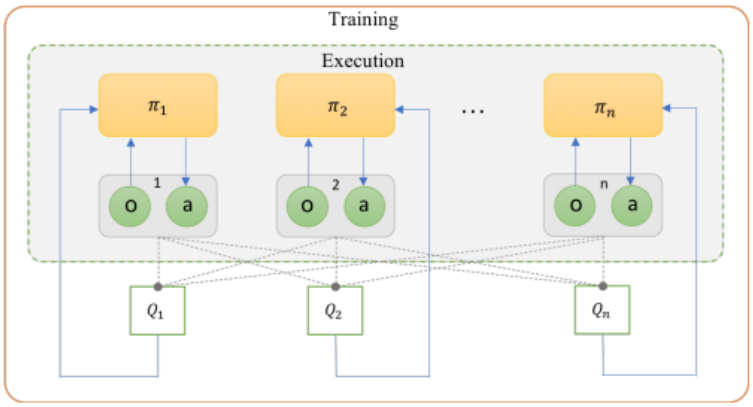
Τα συστήματα πολλαπλών πρακτόρων (MASs) έχουν προσελκύσει μεγάλο ενδιαφέρον λόγω της ικανότητάς τους να επιλύουν σύνθετα προβλήματα μέσω της συνεργασίας πολλαπλών πρακτόρων. Σε ένα τέτοιο σύστημα οι πράκτορες επικοινωνούν μεταξύ τους και αλληλεπιδρούν με το περιβάλλον όπως φαίνεται στο Σχήμα 2.7. Το αντίστοιχο MDP πρόβλημα μπορεί να γενικευτεί σε ένα στοχαστικό παίγνιο ή παίγνιο Markov (Markov game). Αν n είναι το πλήθος των πρακτόρων τότε το παιχνίδι Markov ορίζεται από τα εξής στοιχεία:

- Ένα σύνολο καταστάσεων S.
- Το σύνολο κοινών ενεργειών (joint actions) $A = A_1 \times A_2 \times \cdots \times A_n$ όπου A_i, $i = 1, 2, \ldots, n$ είναι το σύνολο των επιτρεπτών ενεργειών κάθε πράκτορα.
- Τις πιθανότητες μεταβάσεων που περιγράφονται ως $p : S \times A \times S \to [0, 1]$.
- Τη συνάρτηση ανταμοιβής $R : S \times A \times S \to \mathbb{R}^n$ ή ισοδύναμα τη συνάρτηση ανταμοιβής κάθε πράκτορα $R_i : S \times A \times S \to \mathbb{R}$.

Κάθε πράκτορας εκτελεί ενέργειες με στόχο να μεγιστοποιήσει το $\mathbb{E}[G_{i,t}|\pi] = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{i,t+k+1}|\pi]$. Η value function χαρακτηρίζεται ως $Q^\pi : S \times A \to \mathbb{R}^n$.

Η περιοχή των συστημάτων πολλαπλών πρακτόρων είναι μία αρκετά ανοιχτή ερευνητική περιοχή χωρίς αυστηρά μαθηματικά θεμέλια αλλά πειραματικά δίνει καλά αποτελέσματα. Μία βασική πρόκληση που προκύπτει είναι η διαχείριση της ετερογένειας που ενδέχεται να εμφανίζουν οι πράκτορες αλλά και ο τρόπος με τον οποίο μοντελοποιούνται κατάλληλα συλλογικοί ή ανταγωνιστικοί στόχοι. Σε αντίθεση με συστήματα ενός πράκτορα (single-agent), κάθε πράκτορας δεν παρατηρεί μόνο το αποτέλεσμα της δικιάς του ενέργειας αλλά και την συμπεριφορά όλων των υπολοίπων πρακτόρων. Η μάθηση μεταξύ πολλαπλών πρακτόρων είναι αρκετά πιο περίπλοκη διαδικασία εφόσον κάποια αλλαγή στην πολιτική ενός πράκτορα ενδέχεται να

Σχήμα 2.8: Κεντρική εκπαίδευση και αποκεντρικοποιημένη εκτέλεση πολλαπλών πρακτόρων [NNN19]



επηρεάσει την βέλτιστη πολιτική άλλου πράκτορα. Ένα άλλο πρόβλημα που προκύπτει σε πρακτικά προβλήματα είναι ότι ενδέχεται κάποιοι πράκτορες να μην μπορούν να παρατηρήσουν πλήρως την κατάσταση του περιβάλλοντος. Αυτά τα προβλήματα είναι γνωστά ως Μερικώς Παρατηρήσιμες Διαδικασίες Απόφασης Markov (Partially Observable Markov Decision Processes).

2.3.6 Εκπαίδευση σε Συστήματα Πολλαπλών Πρακτόρων

Άμεση επέκταση των single-agent συστημάτων αποτελούν τα multi-agent συστήματα, όπου ο κάθε πράκτορας μαθαίνει ανεξάρτητα θεωρώντας όλους τους υπόλοιπους πράκτορες μέρος του περιβάλλοντος όπως προτάθηκε στην εργασία [Tam+17]. Όμως, η μέθοδος αυτή είναι αρκετά επιρρεπής στην υπερπροσαρμογή (overfitting) [Lan+17], δηλαδή η μάθηση προσαρμόζεται υπερβολικά στα δείγματα εκπαίδευσης και το εκπαιδευμένο μοντέλο δεν έχει καλή γενίκευση σε τυχαία δείγματα. Μία εναλλακτική και αρκετά δημοφιλής προσέγγιση είναι η *κεντρική εκπαίδευση* (*centralized training*) και η *αποκεντρωμένη εκτέλεση* (*decentralized execution*), όπου ένα σύνολο πρακτόρων μπορεί να εκπαιδευτεί ταυτοχρόνως μέσω ενός καναλιού επικοινωνίας [KB16]. Στις αποκεντρικοποιημένες πολιτικές οι πράκτορες επιλέγουν ενέργειες βασισμένες σε τοπικές παρατηρήσεις, δηλαδή οι πράκτορες μπορεί να βρίσκονται σε διαφορετικές καταστάσεις και με βάση αυτές να επιλέγουν τις αντίστοιχες ενέργειες. Τέτοιες πολιτικές έχουν πλεονεκτήματα σε περιβάλλοντα όπου οι πράκτορες έχουν μερική γνώση του περιβάλλοντος. Η κεντρική εκπαίδευση αποκεντρωμένων πολιτικών έχει εφαρμοστεί με επιτυχία σε αρκετές δημοσιεύσεις [Foe+18].

Η κεντρική εκπαίδευση έχει αρκετές διαφορετικές προσεγγίσεις. Μία από τις πιο γνωστές είναι ο διαμοιρασμός παραμέτρων (parameter sharing). Σε αυτήν την προσσέγιση η εκπαίδευση των πρακτόρων γίνεται συλλέγοντας εμπειρίες (δείγματα εκπαίδευσης) από όλους τους πράκτορες. Σε κάθε βήμα ο πράκτορας αποκτά τη δική του εμπειρία της μορφής $(s_i^{(t)}, a_i^{(t)}, r_i^{(t+1)}, s_i^{(t+1)})$, όπου η $s_i^{(t)}$ είναι η κατάσταση που παρατηρεί ο πράκτορας i, $a_i^{(t)}$ είναι η ενέργεια που εκτε-

λεί, $r_i^{(t+1)}$ το σήμα ανταμοιβής που λαμβάνει και $s_i^{(t+1)}$ η κατάσταση στην οποία μεταβαίνει [GEK17].

Στο Σχήμα 2.8 παρουσιάζεται η αποκεντρωμένη εκτέλεση στους πράκτορες. Ωστόσο, η εκπαίδευση των πρακτόρων γίνεται ταυτόχρονα. Υποθέτωντας διαμοιρασμό παραμέτρων, το μοντέλο ενός πράκτορα επεκτείνεται σε σύστημα πολλαπλών πρακτόρων.

3

Μοντελοποίηση Προβλήματος

Στο παρόν κεφάλαιο, παρουσιάζεται το σύστημα, βάσει του οποίου επιλύεται το πρόβλημα ββελτιστοποίησης της παρούσας διπλωματικής εργασίας. Παράλληλα, αναπτύσσεται και περιγράφεται αναλυτικά τόσο το επικοινωνιακό μοντέλο όσο και το μοντέλο του Jake για την μαθηματική περιγραφή του κέρδους καναλιού.

3.1 Περιγραφή Συστήματος Μοντελοποίησης

Θεωρούμε ένα downlink σύστημα μιας κυκλικής κυψέλης, αποτελούμενη από έναν BS στο κέντρο αυτής και $|N|$ χρήστες, όπου $N = \{1, \ldots, n, \ldots, |N|\}$ το σύνολο των χρηστών. Ένα παράδειγμα τέτοιου συστήματος με $|N| = 4$ απεικονίζεται στο Σχήμα 3.1. Στο RSMA το μήνυμα W_n που προορίζεται για το χρήστη n διαρείται σε δύο μηνύματα $W_{c,n}$ και $W_{p,n}$. Το $W_{c,n}$ ονομάζεται κοινό μέρος και το $W_{p,n}$ ιδιωτικό μέρος του μηνύματος W_n [Cle+19]. Τα κοινά μέρη $\{W_{c,1},\ldots,W_{c,|N|}\}$ συνδυάζονται σε ένα κοινό μήνυμα W_C το οποίο στη συνέχεια κωδικοποιείται σε ένα κοινό stream s_0. Το s_0 αποκωδικοποιείται από όλους τους χρήστες. Το ιδιωτικό μέρος $W_{p,n}$ κάθε χρήστη κωδικοποιείται σε ένα ανεξάρτητο ιδιωτικό stream s_n το οποίο αποκωδικοποιεί μόνο ο αντίστοιχος χρήστης.

Το σήμα που μεταδίδει ο BS μπορεί να εκφραστεί ως:

$$x = \sqrt{p_0}s_0 + \sum_{n=1}^{|N|} \sqrt{p_n}s_n, \tag{3.1}$$

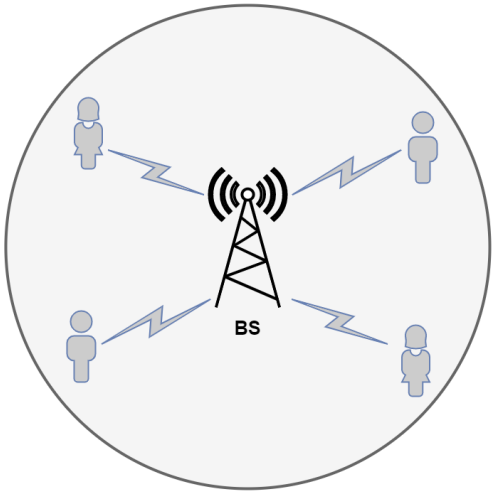
όπου p_0 είναι η ισχύς εκπομπής του s_0 και p_n η ισχύς εκπομπής του s_n.

Το σήμα που λαμβάνει ο χρήστης n εκφράζεται ως:

$$y_n = \sqrt{g_n}x + n_n = \sqrt{g_n p_0}s_0 + \sum_{j=1}^{|N|} \sqrt{g_n p_j}s_j + n_n, \quad \forall n \in N, \tag{3.2}$$

όπου g_n είναι το κέρδος καναλιού της ασύρματης ζεύξης μεταξύ BS και χρήστη n και n_n είναι ο Προσθετικός Λευκός Θόρυβος Gauss (Additive White Gaussian Noise-AWGN) της αντίστοιχης ζεύξης με διασπορά σ^2. Κάθε χρήστης αποκωδικοποιεί πρώτα το κοινό stream αντιμετωπίζοντας την παρεμβολή από τα ιδιωτικά streams ως θόρυβο. Επομένως, ο ρυθμός

Σχήμα 3.1: Παράδειγμα κυκλικής κυψέλης με 4 χρήστες



με τον οποίο μπορεί ο χρήστης n να αποκωδικοποιήσει το s_0 με κανονικοποίηση ως προς το εύρος ζώνης είναι:

$$r_n^c = \log_2\left(\frac{g_n p_0}{g_n \sum_{j=1}^{|N|} p_j + \sigma^2}\right), \quad \forall n \in N, \tag{3.3}$$

Χωρίς βλάβη της γενικότητας μπορούμε να υποθέσουμε ότι $g_1 \leq g_2 \cdots \leq g_N$. Ο κοινός ρυθμός αποκωδικοποίησης r^c είναι:

$$\begin{aligned} r^c &= \min_n r_n^c \\ &= \min_n \log_2\left(\frac{g_n p_0}{g_n \sum_{j=1}^{|N|} p_j + \sigma^2}\right) \\ &= \log_2\left(\frac{g_1 p_0}{g_1 \sum_{j=1}^{|N|} p_j + \sigma^2}\right) \\ &= r_1^c. \end{aligned} \tag{3.4}$$

Για να επιτευχθεί αποκωδικοποίηση του κοινού stream από τον δέκτη κάθε χρήστη πρέπει να ισχύει ο ακόλουθος περιορισμός:

$$g_n p_0 - g_n \sum_{j=1}^{|N|} p_j - \sigma^2 \geq p_{tol}, \quad \forall n \in N, \tag{3.5}$$

όπου p_{tol} είναι η ελάχιστη διαφορά ισχύος στον δέκτη που πρέπει να έχει το s_0 με τον συνολικό θόρυβο (ιδιωτικά streams και θόρυβος καναλιού) ώστε να είναι δυνατή η ορθή ανάκτηση του s_0. Η τιμή p_{tol} εξαρτάται από την κατασκευή του δέκτη. Η σχέση (3.5) γίνεται:

$$p_0 - \sum_{j=1}^{|N|} p_j \geq \frac{p_{tol} + \sigma^2}{g_1} \geq \frac{p_{tol} + \sigma^2}{g_n}, \quad \forall n \in N. \tag{3.6}$$

Δεδομένου του κοινού ρυθμού αποκωδικοποίησης r^c ο BS επιλέγει τους ρυθμούς μετάδοσης $\{c_1, c_2, \ldots, c_{|N|}\}$ του κοινού stream προς τον αντίστοιχο χρήστη υπό τον περιορισμό:

$$\sum_{j=1}^{|N|} c_j \leq r^c. \tag{3.7}$$

Αφού ο χρήστης αποκωδικοποιήσει το s_0 στη συνέχεια αποκωδικοποιεί το ιδιωτικό stream που αντιστοιχεί στον ίδιο αντιμετωπίζοντας τα υπόλοιπα ιδιωτικά streams ως θόρυβο. Επομένως, ο ρυθμός μετάδοσης των ιδιωτικών streams είναι:

$$r_n^p = \log_2\left(1 + \frac{g_n p_n}{g_n \sum_{j=1\neq n}^{|N|} p_j + \sigma^2}\right), \quad \forall n \epsilon N. \tag{3.8}$$

Καταληκτικά, ο συνολικός ρυθμός μετάδοσης για τον χρήστη n είναι:

$$R_n = c_n + r_n^p = c_n + \log_2\left(1 + \frac{g_n p_n}{g_n \sum_{j=1\neq n}^{|N|} p_j + \sigma^2}\right), \quad \forall n \epsilon N. \tag{3.9}$$

3.2 Περιγραφή Κέρδους Καναλιού

Στην παρούσα διπλωματική θεωρούμε ότι η χρονική εξέλιξη ενός τηλεπικοινωνιακού συστήματος, στην περίπτωση μας μίας κυκλικής κυψέλης με 1 BS και $|N|$ χρήστες, πραγματοποιείται σε χρονικές σχισμές (time slots). Σε κάθε χρονική σχισμή το σύστημα μένει σταθερό. Η μεταβολή των χαρακτηριστικών του συστήματος πραγματοποιείται στην έναρξη της επόμενης χρονικής σχισμής.

Ένα από τα βασικά χαρακτηριστικά του τηλεπικοινωνιακού περιβάλλοντος που καλούμαστε να προσομοιώσουμε είναι τα κέρδη των ασύρματων ζεύξεων μεταξύ BS και χρηστών. Υιοθετούμε ένα block fading μοντέλο, δηλαδή ένα μοντέλο όπου η διαδικασία εξασθένισης είναι περίπου σταθερή για αρκετές χρονικές σχισμές. Το κέρδος μεταξύ BS και χρήστη n κατά τη διάρκεια της χρονικής σχισμής t συμβολίζεται με $g_n^{(t)}$ και ορίζεται ως:

$$g_n^{(t)} = |h_n^{(t)}|^2 \alpha_n^{(t)}, \quad \forall n \in N. \tag{3.10}$$

Ο όρος $\alpha_n^{(t)}$ περιγράφει της διαλείψεις μεγάλης κλίμακας και περιλαμβάνει τις απώλειες διαδρομής και τις απώλειες σκίασης. Οι απώλειες αυτές είναι σταθερές για αρκετές διαδοχικές χρονικές σχισμές. Όπως θα δούμε και στη συνέχεια, το σύνολο των χρονικών σχισμών που οι διαλείψεις μεγάλης κλίμακας δεν μεταβάλλονται συνιστά ένα επεισόδιο. Σύμφωνα με το πρότυπο LTE [v80], οι απώλειες διαδρομής μεταξύ BS και χρηστών είναι

$$PL(d_n) = 120.9 + 37.6 \log_{10}(d_n)(dB), \quad \forall n \in N, \tag{3.11}$$

όπου d_n η απόσταση του χρήστη n από τον BS σε km. Οι απώλειες σκίασης περιγράφονται από μία λογαριθμοκανονική κατανομή μηδενικής μέσης τιμής και διασποράς σ_2^2.

Ο όρος $h_n^{(t)}$ περιγράφει τις διαλείψεις μικρής κλίμακας Rayleigh. Ο όρος αυτός μεταβάλλεται ανά χρονική σχισμή και περιγράφεται ως μία πρώτης τάξεως Gauss-Markov διαδικασία:

$$h_n^{(t)} = \rho h_n^{(t-1)} + \sqrt{1-\rho^2} e_n^{(t)}, \quad \forall t \geq 1 \quad \kappa\alpha\iota \quad \forall n \in N, \tag{3.12}$$

όπου $e_n^{(t)}$ είναι μία Κυκλικά Συμμετρική Σύνθετη Γκαουσιανή κατανομή (Circularly Symmetric Complex Gaussian-CSCG) με μοναδιαία διασπορά. Οι κατανομές $e_n^{(1)}, e_n^{(2)}, \ldots$ είναι ανεξάρτητες μεταξύ τους. Παρατηρούμε ότι ο όρος $\rho h_n^{(t-1)}$ μοντελοποιεί την εξάρτηση του κέρδους την χρονική σχισμή t από το κέρδος την χρονική σχισμή $t-1$, ενώ ο όρος $\sqrt{1-\rho^2} e_n^{(t)}$ μοντελοποιεί την τυχαιότητα του κέρδους. Ο όρος ρ ονομάζεται συσχέτιση και ορίζεται ως:

$$\rho = J_0(2\pi f_d T), \tag{3.13}$$

όπου J_0 είναι η μηδενικής τάξεως συνάρτηση Bessel πρώτου είδους, f_d είναι η μέγιστη συχνότητα Doppler των χρηστών και T η χρονική διάρκεια της σχισμής.

3.3 Διατύπωση Προβλήματος

Ο στόχος του προβλήματος βελτιστοποίησης είναι να μεγιστοποιήσουμε την ενεργειακή απόδοση του BS. Επομένως, το πρόβλημα βελτιστοποίησης διατυπώνεται ως:

$$\max_{\boldsymbol{c},\boldsymbol{P},p_0} EE = \frac{1}{|N|} \frac{\sum_{n=1}^{|N|} w_n R_n}{\sum_{n=1}^{|N|} p_n + p_0} \tag{3.14}$$

$$\text{s.t.} \quad \sum_{n=1}^{|N|} c_n \leq r_1^c, \tag{3.15}$$

$$p_0 - \sum_{j=1}^{|N|} p_j \geq \frac{p_{tol} + \sigma^2}{g_1}, \tag{3.16}$$

$$\sum_{n=1}^{|N|} p_n + p_0 \leq P_{max}, \tag{3.17}$$

$$c_n, p_n \geq 0, \quad \forall n \quad \kappa\alpha\iota \quad p_0 \geq 0, \tag{3.18}$$

όπου $\boldsymbol{c} = \begin{bmatrix} c_1 & c_2 & \ldots & c_{|N|} \end{bmatrix}^T$ το διάνυσμα κοινών ρυθμών, $\boldsymbol{P} = \begin{bmatrix} p_1 & p_2 & \ldots & p_{|N|} \end{bmatrix}^T$ το διάνυσμα ισχύος των ιδιωτικών streams και p_0 η ισχύς του κοινού stream.

Ο όρος EE της εξίσωσης (3.14) εκφράζει την Ενεργειακή Απόδοση του δικτύου και ορίζεται ως το σταθμισμένο άθροισμα των ρυθμών μετάδοσης (WSR) ανά μονάδα ισχύος εκπομπής ανά χρήστη και μετριέται σε bits/J/Hz. Ο περιορισμός (3.15) εξασφαλίζει ότι όλοι οι χρήστες μπορούν να αποκωδικοποιήσουν το κοινό stream, η σχέση (3.16) περιγράφει τον περιορισμό εκείνο που εξασφαλίζει ότι ο δέκτης μπορεί να διαχωρίσει ορθά το κοινό μήνυμα από τα ιδιωτικά μηνύματα, ενώ ο περιορισμός (3.17) εκφράζει ένα άνω όριο στην μέγιστη συνολική ισχύ που μπορεί να μεταδώσει ο BS.

Τέλος, εφόσον μελετάμε ένα δίκτυο πραγματικού χρόνου στο οποίο τα κέρδη των καναλιών μεταβάλλονται ανά χρονική σχισμή, προκύπτει ότι το πρόβλημα βελτιστοποίησης που παρουσιάστηκε αλλάζει σε κάθε χρονική σχισμή.

4

Περιγραφή Μεθόδου Επίλυσης

Σε αυτό το κεφάλαιο θα παρουσιαστούν αναλυτικά οι αλγόριθμοι επίλυσης του προβλήματος βελτιστοποίησης που διατυπώθηκε στην Ενότητα 3.3. Αρχικά θα διατυπώσουμε τα στοιχεία του MDP μοντέλου όπως αυτά ορίστηκαν στην Ενότητα 2.3.2. Έπειτα, θα παρουσιάσουμε τους αλγορίθμους DQL και REINFORCE που επιστρατεύονται προς επίλυση του προβλήματος, ενώ θα προηγηθεί μία σύντομη εισαγωγή στη Βαθιά Ενισχυτική Μάθηση (Deep Reinforcement Learning,DRL). Στη συνέχεια, θα προχωρήσουμε σε μία τροποποίηση του MDP πλαισίου ώστε να επιλύσουμε ένα εναλλακτικό πρόβλημα βελτιστοποίησης που στοχεύει στη μεγιστοποίηση του WSR χωρίς να λαμβάνει υπόψην την ενεργειακή αποδοτικότητα, το οποίο και θα χρησιμοποιηθεί για την αξιολόγηση της μεγιστοποίησης του EE ως στόχος βελτιστοποίησης. Τέλος, θα περιγραφεί ο αλγόριθμος Σταθμισμένου Μέσου Τετραγωνικού Σφάλματος (Weighted Mean Square Error-WMSE) ο οποίος χρησιμοποιείται αρκετά συχνά σε προβλήματα ανάθεσης ισχύος ώστε να έχουμε ένα μέτρο σύγκρισης για την αξιολόγηση των αλγορίθμων ενισχυτικής μάθησης.

4.1 Μοντελοποίηση Προβλήματος ως MDP

Υπάρχουν δύο βασικές κατηγορίες προβλημάτων ενισχυτικής μάθησης. Η πρώτη είναι τα βασισμένα στο μοντέλο προβλήματα (model-based) στα οποία η δυναμική του περιβάλλοντος (πιθανότητες μεταβάσεων) είναι γνωστή στον πράκτορα. Η κατηγορία αυτή είναι μία ευρύτερη περιοχή προβλημάτων που περιλαμβάνει και προβλήματα δυναμικού προγραμματισμού (dynamic programming) και δε βασίζεται τόσο στη μάθηση όσο στον σχεδιασμό πολιτικών που να προσαρμόζονται όσο το δυνατόν πιο βέλτιστα στο γνωστό περιβάλλον. Η δεύτερη κατηγορία είναι τα ελεύθερα από το περιβάλλον (model-free) προβλήματα στα οποία δεν υπάρχει γνώση της δυναμικής του περιβάλλοντος και ο πράκτορας μαθαίνει τη βέλτιστη πολιτική παρατηρώντας τις επιπτώσεις που έχουν οι ενέργειές του στο περιβάλλον. Επομένως, σε αυτά τα προβλήματα δεν μοντελοποιούνται οι κατανομές μετάβασεων 2.3.2.

Οι αλγόριθμοι επίλυσης που υλοποιούμε στην παρούσα εργασία είναι model-free, δηλαδή για να ορίσουμε το MDP πρόβλημα αρκεί να ορίσουμε το σύνολο των καταστάσεων, το σύνολο ενεργειών και τη συνάρτηση ανταμοιβής.

Στην ενότητα αυτή, θα ορίσουμε το περιβάλλον του προβλήματος και στη συνέχεια τις καταστάσεις, τις ενέργειες και την συνάρτηση ανταμοιβής.

4.1.1 Περιβάλλον Πολλαπλών Πρακτόρων

Το περιβάλλον απότελείται από μία κυκλική κυψέλη με έναν BS στο κέντρο της και ένα σύνολο $N = \{1, \ldots, n, \ldots, |N|\}$ χρηστών εντός αυτής, όπως απεικονίζεται στο Σχήμα 3.1. Οι χρήστες βρίσκονται σε τυχαίες θέσεις γύρω από τον BS. Η ελάχιστη απόσταση χρήστη από το κέντρο της κυψέλης είναι R_{min} ενώ η μέγιστη R_{max}. Το περιβάλλον έχει διάρκεια στο χρόνο. Συγκεκριμένα, ο χρόνος διαιρείται σε χρονικές σχισμές διάρκειας T. Σε κάθε χρονική σχισμή, το περιβάλλον μένει σταθερό και στην έναρξη της επόμενης σχισμής αλλάζει. Όπως εξηγήθηκε και στην Ενότητα 3.2, το βασικό χαρακτηριστικό που μεταβάλλεται είναι τα κέρδη των καναλιών μεταξύ BS και χρηστών.

Ακολουθούμε μία προσέγγιση πολλαπλών πρακτόρων με τα χαρακτηριστικά που παρουσιάστηκαν στις Ενότητες 2.3.5 και 2.3.6. Συγκεκριμένα, υιοθετούμε μία προσέγγιση κεντρικής εκπαίδευσης και αποκεντρωμένης εκτέλεσης, στην οποία οι πράκτορες παρατηρούν μια μοναδική τοπική κατάσταση και επιλέγουν κατάλληλη ενέργεια με βάση αυτή. Η εκπαίδευση, όμως, των πρακτόρων γίνεται με δείγματα εμπειριών που συλλέγονται από όλους τους πράκτορες.

Όπως και στην εργασία [Ben09] κάθε ιδιωτικό stream θεωρείται ένας ξεχωριστός πράκτορας. Παρόμοια με την [HW98] εργασία, συμβολίζουμε την κατάσταση του πράκτορα i ως $s_i \in S_i$, όπου S_i είναι το σύνολο όλων των καταστάσεων που μπορεί να βρεθεί ο πράκτορας i. Η κατάσταση s_i, όπως θα δούμε και στη συνέχεια, αποτελείται από χαρακτηριστικά του περιβάλλοντος που αφορούν το stream s_i.

Στο τέλος της χρονικής σχισμής t ο πράκτορας i επιλέγει μία ενέργεια $a_i^{(t)}$ ως συνάρτηση της κατάστασης $s_i^{(t)}$ με βάση την τωρινή πολιτική π. Όλοι οι πράκτορες είναι συγχρονισμένοι και αποφασίζουν ταυτόχρονα, όμως κανείς δεν γνωρίζει τις ενέργειες των άλλων. Επειδή, όμως, οι ενέργειες που επιλέγουν επηρεάζουν το κοινό περιβάλλον και επομένως τις καταστάσεις όλων των πρακτόρων, κάθε πράκτορας με βάση προηγούμενες εμπειρίες του μπορεί να εκτιμήσει την επίδραση που έχει μία ενέργεια $a_i^{(t)}$ στις μελλοντικές αποφάσεις όλων των πρακτόρων.

Τέλος, σημειώνουμε ότι θεωρούμε το πρόβλημα μας episodic. Ανά N_s χρονικές σχισμές οι πράκτορες εισέρχονται σε τερματική κατάσταση, η διεπαφή περιβάλλον-πρακτόρων «σπάει» και το περιβάλλον ξεκινάει σε τυχαία αρχική κατάσταση. Οι θέσεις των χρηστών, όπως και τα κέρδη των καναλιών, αρχικοποιούνται σε νέες τιμές στην αρχή κάθε επεισοδίου.

4.1.2 Καταστάσεις, Ενέργειες και Ανταμοιβή

Καταστάσεις

Οργανώνουμε την πληροφορία που παρατηρεί ο πράκτορας i σε 8 χαρακτηριστικά. Τα χαρακτηριστικά αυτά περιέχουν πληροφορία που αφορά την μετάδοση του stream s_i στην

ασύρματη ζεύξη BS-χρήστη i. Τα 8 αυτά χαρακτηριστικά συνθέτουν την τοπική κατάσταση $s_i^{(t)}$ και είναι τα ακόλουθα:

(i) Το κέρδος του καναλιού $g_i^{(t)}$ τη χρονική σχισμή t.

(ii) Την παρεμβολή, που προκύπτει από τα υπόλοιπα ιδιωτικά streams και τον θόρυβο καναλιού, στο κανάλι BS-χρήστη i την χρονική σχισμή t. Η παρεμβολή αυτή είναι $\sum_{j\in N, j\neq i} g_i^{(t)} p_j^{(t-1)} + \sigma^2$. Σημειώνουμε ότι η ισχύς κοινού και ιδιωτικών streams αλλάζει στο τέλος κάθε χρονικής σχισμής και μένει σταθερή κατά τη διάρκεια της επόμενης. Επομένως κατά τη διάρκεια της χρονικής σχισμής t οι ισχύς των ιδιωτικών streams είναι $\boldsymbol{P^{(t-1)}} = \begin{bmatrix} p_1^{(t-1)} & p_2^{(t-1)} & \dots & p_{|N|}^{(t-1)} \end{bmatrix}$ ενώ η ισχύς του κοινού stream είναι $p_0^{(t-1)}$.

(iii) Το κέρδος του καναλιού $g_i^{(t-1)}$ τη σχισμή $t-1$

(iv) Την παρεμβολή, που προκύπτει από τα υπόλοιπα ιδιωτικά streams και τον θόρυβο καναλιού, στο κανάλι BS-χρήστη i, την χρονική σχισμή $t-1$. Η παρεμβολή αυτή είναι $\sum_{j\in N, j\neq i} g_i^{(t-1)} p_j^{(t-2)} + \sigma^2$.

(v) Η ισχύς $p_i^{(t-1)}$ του stream s_i.

(vi) Η ισχύς $p_0^{(t-1)}$ του ιδιωτικού stream s_0.

(vii) Ο ρυθμός μετάδοσης του ιδιωτικού stream, $r_i^{(t)}$.

(viii) Ο ρυθμός μετάδοσης του κοινού stream για τον χρήστη i, $c_i^{(t)}$.

Ο λόγος που επιλέξαμε να μη θεωρήσουμε έναν επιπλέον πράκτορα για το κοινό stream s_0 είναι ότι επιθυμούμε οι πράκτορες να είναι **ομογενείς**, ώστε να είναι εφικτός ο διαμοιρασμός παραμέτρων (parameter sharing). Οι πράκτορες εκτελούν ακριβώς την ίδια λειτουργία αλλά για διαφορετική «περιοχή» stream-ζεύξη. Για το λόγο αυτό, οι καταστάσεις $s_i^{(t)}$ για $i = 0, 1, 2, ..$ έχουν ακριβώς τα ίδια χαρακτηριστικά, αλλά οι τιμές των χαρακτηριστικών αλλάζουν ανάλογα με την «περιοχή» που παρατηρεί ο κάθε πράκτορας.

Παρατηρούμε ότι μέσω της εισαγωγής των χαρακτηριστικών (vi) και (viii) οι πράκτορες αποκτούν γνώση για το κοινό stream. Το χαρακτηριστικό (vi) ενημερώνει τον πράκτορα για την ισχύ του κοινού stream και είναι ίδιο για όλους τους πράκτορες, ενώ το χαρακτηριστικό (viii) ενημερώνει τον πράκτορα i για τον κοινό ρυθμό μετάδοσης που ανατίθεται στο χρήστη i. Επίσης, τα χαρακτηριστικά (iii) και (vi) είναι ίδια με τα (i) και (ii) για δύο διαδοχικές χρονικές σχισμές. Αυτός είναι ένας τρόπος να μεταφέρουμε άμεσα κάποιες παρατηρήσεις σε επόμενες χρονικές σχισμές. Είναι δηλαδή σαν ένα είδος **μνήμης** όπου ο πράκτορας αποθηκεύει το κέρδος καναλιού και την παρεμβολή ώστε οι παρατηρήσεις αυτές να είναι διαθέσιμες στην επόμενη χρονική σχισμή. Θα μπορούσε κάποιος να θεωρήσει περιπτώσεις όπου μεταφέρονται M διαδοχικά στιγμιότυπα των χαρακτηριστικών (i) και (ii) στην κατάσταση $s_i^{(t)}$ από τις προηγούμενες $t-M$ σχισμές, ώστε να υπάρχει περισσότερη μεταφορά γνώσης. Αυτό όμως κοστίζει σε πολυπλοκότητα διότι ο χώρος καταστάσεων αποκτά μεγαλύτερη διάσταση. Αυτό

σημαίνει ότι οι μέθοδοι προσέγγισης συνάρτησης τιμής (value function approximators), που εξηγήθηκαν στην Ενότητα 2.3.3, απαιτούν πιο πολύπλοκες συναρτήσεις τιμής για να αναπαραστήσουν είτε την state-action value είτε απευθείας την πολιτική. Επίσης, όσο αυξάνεται η διάσταση του χώρου καταστάσεων το gradient descent που εξηγήθηκε στην Ενότητα 2.3.4 γίνεται υπολογιστικά πιο ακριβό. Για αυτό το λόγο, η επιλογή των χαρακτηριστικών πρέπει να γίνει προσεκτικά και να συμπεριληφθούν όσα χαρακτηριστικά είναι απαραίτητα.

Παρατηρούμε επίσης ότι ο χώρος καταστάσεων είναι **συνεχής**, αφού όλα τα χαρακτηριστικά του παίρνουν τιμές σε συνεχή διαστήματα. Αυτό είναι σημαντικό διότι οι τεχνικές Βαθιάς Ενισχυτικής Μάθησης που θα εξετάσουμε στη συνέχεια έχουν εφαρμογή σε συνεχείς χώρους.

Συμπερασματικά, η κατάσταση κάθε πράκτορα περιγράφεται από το παρακάτω διάνυσμα 8 χαρακτηριστικών:

$$s_i^{(t)} = \left[g_i^{(t)}, \sum_{j\in N, j\neq i} g_i^{(t)} p_j^{(t-1)} + \sigma^2, g_i^{(t-1)}, \sum_{j\in N, j\neq i} g_i^{(t-1)} p_j^{(t-2)} + \sigma^2, p_i^{(t-1)},\ p_0^{(t-1)}, r_i^{(t)}, c_i^{(t)} \right]^T, \quad \forall i \in N. \tag{4.1}$$

Ενέργειες

Οι ενέργειες που αποφασίζει ο πράκτορας i συμβολίζονται με $a_i \in A_i$, όπου A_i είναι ο χώρος αποφάσεων του πράκτορα i. Ο χώρος αποφάσεων ορίζεται ως οι στάθμες της ισχύος που μπορούν να χρησιμοποιηθούν για την εκπομπή του ιδιωτικού stream s_i. Δηλαδή $a_i = p_i$. Επειδή, όμως, οι αλγόριθμοι που θα χρησιμοποιήσουμε για την επίλυση εφαρμόζονται για διακριτό χώρο αποφάσεων, απαιτείται κβαντισμός του εύρους της ισχύος σε διακριτές στάθμες.

Έστω $P_{max,i}$ η μέγιστη επιτρεπτή ισχύς του stream s_i. Ορίζουμε την $P_{max,i}$ ως $P_{max,i} = \frac{P_{max}}{|N|+1}$. Επίσης θεωρούμε ότι η ελάχιστη ισχύ εκπομπής του s_i είναι $P_{min,i}$.

Για την εύρεση του $p_0^{(t)}$ ακολουθούμε της εξής διαδικασία: Θεωρούμε το σύνολο $N_{p_0}^{(t)} = \left[\frac{P_{max} - \sum_{j\in N} a_j^{(t)}}{|N_{p_0}|}, \frac{P_{max} - \sum_{j\in N} a_j^{(t)}}{|N_{p_0}|-1}, \ldots, \frac{P_{max} - \sum_{j\in N} a_j^{(t)}}{1} \right]$, όπου $N_{p_0}^{(t)}$ είναι το σύνολο πιθανών τιμών για το $p_0^{(t)}$. Παρατηρούμε ότι για κάθε λύση στο $N_{p_0}^{(t)}$ ικανοποιείται ο περιορισμός (3.17).

Η εξίσωση (3.14) μπορεί γραφτεί ως:

$$\max_{\boldsymbol{c},\boldsymbol{P},p_0} EE = \frac{1}{|N|} \frac{\sum_{n=1}^{|N|} w_n R_n}{\sum_{n=1}^{|N|} p_n + p_0} = \frac{1}{|N|} \frac{\sum_{n=1}^{|N|} w_n c_n}{\sum_{n=1}^{|N|} p_n + p_0} + \frac{1}{|N|} \frac{\sum_{n=1}^{|N|} w_n r_n^p}{\sum_{n=1}^{|N|} p_n + p_0}. \tag{4.2}$$

Ο μόνος άγνωστος όρος στην εξίσωση (4.2) είναι ο $\max_{\boldsymbol{c}} C = \sum_{n=1}^{|N|} w_n c_n$, τον οποίο επιθυμούμε να μεγιστοποιήσουμε. Επομένως μπορούμε να λύσουμε το ακόλουθο πρόβλημα γραμμικού προγραμματισμού προς εύρεση των κοινών ρυθμών:

$$\max_{\boldsymbol{c}} C = \sum_{n=1}^{|N|} w_n c_n \tag{4.3}$$

$$\text{s.t.} \quad \sum_{n=1}^{|N|} c_n \leq r_1^c, \tag{4.4}$$

$$c_n \geq 0. \tag{4.5}$$

Συνεπώς, καταλήγουμε σε ένα σύνολο υποψήφιων λύσεων $(\boldsymbol{c^{(t)}}, p_0^{(t)})$ για την ενέργεια $\boldsymbol{P^{(t)}}$ των πρακτόρων. Για κάθε λύση $(\boldsymbol{c^{(t)}}, \boldsymbol{P^{(t)}}, P_0^{(t)})$ υπολογίζουμε την ενεργειακή απόδοση μέσω της εξίσωσης (3.14) και κρατάμε τη λύση με τη μεγαλύτερη τιμή.

Όπως αναφέραμε παραπάνω, απαιτείται κβαντισμός του εύρους ισχύος των ιδιωτικών streams. Επιλέγουμε τον ακόλουθο λογαριθμοκανονικό κβαντιστή:

$$A_i = \left\{ 0, P_{min,i}, P_{min} \left(\frac{P_{max,i}}{P_{min,i}} \right)^{\frac{1}{|A_i|-2}}, \ldots, P_{max,i} \right\}, \tag{4.6}$$

όπου P_{min} είναι η ελάχιστη μη μηδενική ισχύς εκπομπής ενός ιδιωτικού stream και $|A_i|$ είναι το πλήθος των επιπέδων ισχύος. Στην παρούσα εργασία θεωρούμε $A_i = A, \quad \forall i \in N$.

Ανταμοιβή

Ο σκοπός της συνάρτησης ανταμοιβής όπως τη μοντελοποιούμε στην παρούσα εργασία είναι να ενημερώνει τους πράκτορες για την επίδραση που έχουν οι ενέργειες τους, στο EE του δικτύου. Δηλαδή, όταν η λύση $(\boldsymbol{c}, \boldsymbol{P}, p_0)$ αυξάνει τον όρο $\frac{1}{|N|} \frac{\sum_{n=1}^{|N|} w_n R_n}{\sum_{n=1}^{|N|} p_n + p_0}$ οι πράκτορες μέσω της συνάρτησης ανταμοιβής πρέπει να ενημερώνονται θετικά.

Στην παρούσα εργασία μοντελοποιούμε τη συνάρτηση ανταμοιβής με δύο διαφορετικούς τρόπους, όπως εξηγείται αναλυτικά παρακάτω. Στον έναν τρόπο θεωρούμε ότι η ανταμοιβή διαφέρει σε κάθε πράκτορα, ενώ στον δεύτερο τρόπο οι πράκτορες λαμβάνουν το ίδιο σήμα ανταμοιβής.

Όσον αφορά των πρώτο τρόπο υπολογίζονται τα παρακάτω μεγέθη:

$$C_{k/i}^{(t)} = \log_2 \left(1 + \frac{g_k^{(t)} p_k^{(t-1)}}{g_k^{(t)} \sum_{j \neq i,k} p_j^{(t-1)} + \sigma^2} \right), \tag{4.7}$$

για κάθε διατεταγμενο ζεύγος (k, i) όπου $k \neq i$. Ο όρος $C_{k/i}^{(t)}$ περιγράφει τον ρυθμό μετάδοσης που θα είχε το stream s_k προς τον αντίστοιχο χρήστη εάν δεν υπήρχε η παρεμβολή από το s_i. Στη συνέχεια υπολογίζονται οι όροι:

$$\pi_{i \to k}^{(t)} = w_k \left(C_{k/i}^{(t)} - r_k^{p,(t)} \right). \tag{4.8}$$

Ο όρος $\pi_{i \to k}^{(t)}$ περιγράφει τη μείωση που προκαλεί στον ρυθμό μετάδοσης του s_k η παρεμβολή από το s_i.

Όμοια, μπορούμε να ορίσουμε και τα μεγέθη:

$$C_{0/i}^{(t)} = \log_2\left(1 + \frac{g_1^{(t)} p_k^{(t-1)}}{g_1^{(t)} \sum_{j \neq i} p_j^{(t-1)} + \sigma^2}\right), \quad \forall i \in N, \tag{4.9}$$

και

$$\pi_{i \to 0}^{(t)} = w_k \left(C_{0/i}^{(t)} - r_1^{c,(t)}\right), \quad \forall i \in N, \tag{4.10}$$

όπου $g_1^{(t)} = \min_k g_k^{(t)}$, όπως έχει οριστεί στην Ενότητα 3.1. Το μέγεθος 4.10 περιγράφει τη μείωση του κοινού ρυθμού λόγω της παρεμβολής από το s_i.

Ορίζουμε το σήμα ανταμοιβής του πράκτορα i ως:

$$R_{i,1}^{(t)} = \frac{1}{|N|} \frac{w_i R_i^{(t)} - \beta \sum_{j \neq i} \pi_{i \to j}^{(t)}}{\sum_{n=1}^{|N|} p_n^{(t-1)} + p_0^{(t-1)}} \tag{4.11}$$

όπου ο δείκτης 1 συμβολίζει τη πρώτη μέθοδο. Η ιδέα είναι ότι ο πράκτορας i παίρνει μία θετική ανταμοιβή $w_i R_i$ ανάλογη του ρυθμού μετάδοσης του s_i και μια ποινή (penalty) $\sum_{j \neq i} \pi_{i \to j}^{(t)}$ ανάλογη της μείωσης των ρυθμών κοινού και υπόλοιπων ιδιωτικών streams λόγω της παρεμβολής από το s_i. Η παράμετρος β ρυθμίζει τη βαρύτητα που έχει το penalty στη συνάρτηση ανταμοιβής. Ο παρονομαστής στη συνάρτηση ανταμοιβής είναι η συνολική ισχύς εκπομπής.

Ο δεύτερος τρόπος θεωρεί μια πιο απλή μοντελοποίηση της συνάρτησης ανταμοιβής. Σύμφωνα με αυτήν η ανταμοιβή των πρακτόρων είναι ίδια και ίση με το EE του δικτύου. Δηλαδή:

$$R_{i,2}^{(t)} = \frac{1}{|N|} \frac{\sum_{n=1}^{|N|} w_n R_n^{(t)}}{\sum_{n=1}^{|N|} p_n^{(t-1)} + p_0^{(t-1)}}. \tag{4.12}$$

Μία επιπλέον τροποποίηση που κάνουμε αφορά τον περιορισμό 3.16. Η ιδέα είναι να προσαρτήσουμε τον περιορισμό στη συνάρτηση ανταμοιβής ώστε οι πράκτορες να βρίσκουν ενέργειες που τους ικανοποιούν. Η τελική μορφή της συνάρτηση ανταμοιβής που λαμβάνουν οι πράκτορες είναι:

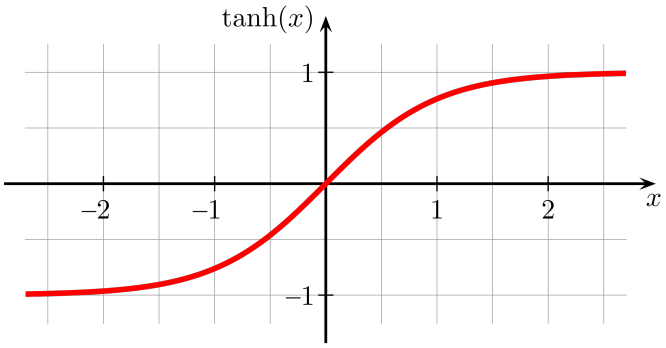
$$r_i^{(t)} = R_{i,j}^{(t)} \quad \text{εάν} \quad p_0^{(t-1)} - \sum_{j=1}^{|N|} p_j^{(t-1)} \geq \frac{p_{tol} + \sigma^2}{g_1^{(t)}}, \tag{4.13}$$

$$r_i^{(t)} = R_{i,j}^{(t)}(1 + tanh(p_0^{(t-1)} - \sum_{j=1}^{|N|} p_j^{(t-1)} - \frac{p_{tol} + \sigma^2}{g_1^{(t-1)}})), \quad \text{εάν} \quad p_0^{(t-1)} - \sum_{j=1}^{|N|} p_j^{(t-1)} \leq \frac{p_{tol} + \sigma^2}{g_1^{(t)}}, \tag{4.14}$$

όπου $i = 1, 2, \ldots, n$ είναι ο δείκτης του πράκτορα και $j = 1, 2$ είναι ο δείκτης της μεθόδου.

Η συνάρτηση $tanh$ έχει το εξής χαρακτηριστικό που εύκολα παρατηρούμε και στο Σχήμα 4.1. Είναι ασυμπτωτική στο $-\infty$ στην τιμή -1. Επομένως, δεδομένου του τρόπου ορισμού της συνάρτησης ανταμοιβής στο (4.14), προκύπτει ότι τείνει στο 0 όσο απομακρυνόμαστε από την ικανοποίηση του περιορισμού 3.16. Αυτό έχει σαν αποτέλεσμα, οι πράκτορες να μαθαίνουν την αρνητική επίδραση της μη ικανοποίησης του περιορισμού.

Σχήμα 4.1: Συνάρτηση tanh.



4.2 Βαθιά Ενισχυτική Μάθηση

Τα τελευταία χρόνια η ανάπτυξη των νευρωνικών δικτύων και συγκεκριμένα των συνελικτικών νευρωνικών δικτύων (Convolutional Neural Networks-CNN) [Ben09] έχει ανοίξει τον δρόμο για την εξερεύνηση νέων τεχνικών στον τομέα της τεχνητής νοημοσύνης. Τα νευρωνικά δίκτυα έχει αποδειχθεί ότι μπορούν να αναπαραστήσουν πολύπλοκα σύνολα δεδομένων αλλά και πολύπλοκες εξαρτήσεις μεταξύ συνόλων. Το πεδίο της βαθιάς ενισχυτικής μάθησης χρησιμοποιεί νευρωνικά δίκτυα ως value function approximators με σκοπό να αναπαραστήσει καλύτερα πιο σύνθετες εξαρτήσεις των συναρτήσεων τιμής από τον χώρο καταστάσεων και αποφάσεων.

Στην Ενότητα 4.1, μοντελοποιήσαμε το πρόβλημα βελτιστοποίησης του Κεφαλαίου 3 ως MDP. Στην ενότητα αυτή, προτείνονται δύο αλγόριθμοι επίλυσης του με χρήση νευρωνικών δικτύων. Στον πρώτο αλγόριθμο, χρησιμοποιούμε ένα νευρωνικό δίκτυο για την αναπαράσταση της action-state value $Q(s, a|\theta)$, ενώ ο δεύτερος αλγόριθμος χρησιμοποιεί ένα νευρωνικό δίκτυο που παραμετροποιεί απευθείας την πολιτική $\pi(a|s, \theta)$, όπου $\theta \in \mathbb{R}^{|\theta|}$ είναι το σύνολο των συναπτικών βαρών του νευρωνικού δικτύου.

4.2.1 Τεχνητά Νευρωνικά Δίκτυα

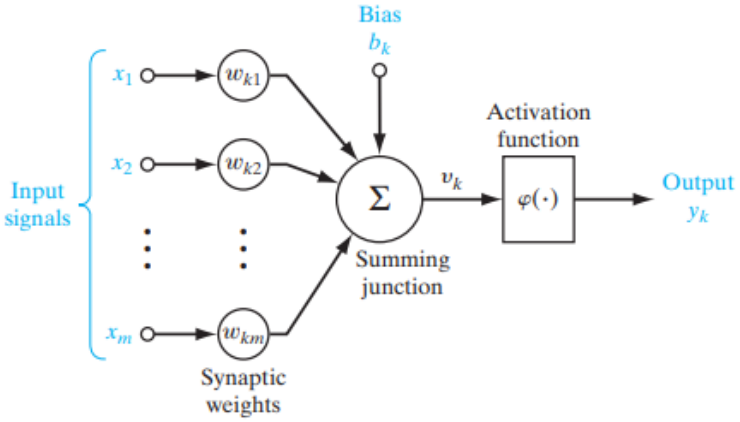
Το βασικό στοιχείο ενός τεχνητού νευρωνικού δικτύου ονομάζεται νευρώνας. Ο νευρώνας είναι μια βασική μονάδα επεξεργασίας που αποτελείται από τα παρακάτω στοιχεία:

(i) Ένα σύνολο συνάψεων, όπου η κάθε σύναψη χαρακτηρίζεται από το βάρος της. Συγκεκριμένα, στην είσοδο της σύναψης j εφαρμόζεται ένα σήμα x_j και διαμέσου της σύναψης πολλαλασιάζεται με το αντίστοιχο βάρος $w_{k,j}$, όπου ο δείκτης k είναι αναγνωριστικό του νευρώνα. Το βάρος μπορεί να παίρνει θετικές ή αρνητικές τιμές.

(ii) Έναν αθροιστή για την άθροιση των σταθμισμένων από τα αντίστοιχα συναπτικά βάρη σημάτων εισόδου. Η λειτουργία αυτή είναι γραμμική.

(iii) Μία σύναρτηση ενεργοποίησης στην έξοδο του νευρώνα για τον περιορισμό του σήματος αλλά και την εισαγωγή μη γραμμικότητας στη σχέση εισόδου-εξόδου.

Σχήμα 4.2: Αρχικτεκτονική Νευρώνα [Hay09].



(iv) Μία εξωτερική πόλωση b_k, που έχει στόχο την αύξηση ή μείωση της δικτυακής διέγερσης.

Μαθηματικά ο νευρώνας μπορεί να περιγραφεί με το παρακάτω ζεύγος εξισώσεων:

$$u_k = \sum_{j=1}^{m} w_{k,j} x_j, \tag{4.15}$$

$$y_k = \phi(u_k + b_k), \tag{4.16}$$

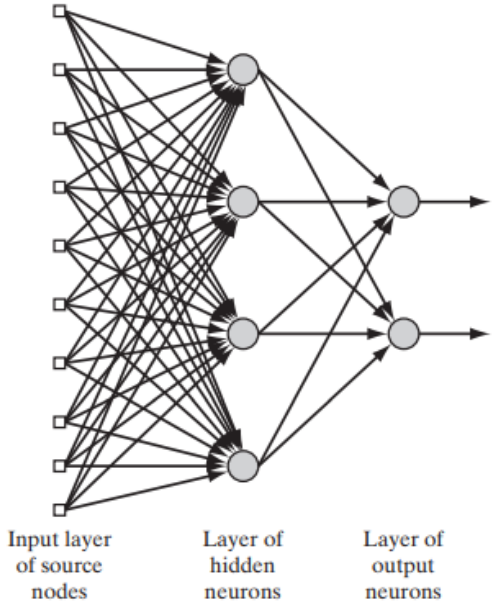
όπου $x_1, x_2, \ldots, x_m$ είναι τα σήματα εισόδου, $w_{k,1}, w_{k,2}, \ldots, w_{k,m}$ είναι τα αντίστοιχα συναπτικά βάρη του νευρώνα k και $\phi()$ η συνάρτηση ενεργοποίησης. Η λειτουργία του νευρώνα απεικονίζεται στο Σχήμα 4.2.

Σε ένα νευρωνικό δίκτυο, οι νευρώνες οργανώνονται σε μορφή επιπέδων. Στην απλούστερη περίπτωση ενός δικτύου έχουμε ένα επίπεδο εισόδου που συνδέεται απευθείας με ένα άλλο επίπεδο νευρώνων εξόδου, αλλά όχι αντίστροφα. Το δίκτυο αυτό ονομάζεται Πρόσθιας Τροφοδότησης. Πιο σύνθετες μορφές περιλαμβάνουν πολλαπλά επίπεδα νευρώνων που ονομάζονται κρυφά επίπεδα. Η αρχιτεκτονική ενός νευρωνικού δικτύου επηρεάζει τον αλγόριθμο μάθησης. Προσθέτοντας περισσότερα κρυφά επίπεδα το νευρωνικό δίκτυο μπορεί να εξαγάγει από την είσοδό του στατιστικά υψηλότερης τάξης. Στο Σχήμα 4.3, απεικονίζεται η αρχιτεκτονική ενός δικτύου Πρόσθιας Τροφοδότησης με ένα κρυφό επίπεδο.

Το νευρωνικό δίκτυο που προτείνουμε είναι Πρόσθιας Τροφοδοτησης με 3 κρυφά επίπεδα με N_1, N_2 και N_3 νευρώνες αντίστοιχα. Το πρώτο επίπεδο είναι το επίπεδο εισόδου με $N_0 = 8$ σήματα εισόδου, ενώ το τελευταίο είναι το επίπεδο εξόδου μεγέθους $N_4 = |A|$.

Εδώ μπορούμε να αντιληφθούμε και το λόγο που επιλέξαμε ένα σύστημα πολλαπλών πρακτόρων. Σε περίπτωση που θεωρούσαμε ένα μόνο πράκτορα, ο χώρος αποφάσεων θα οριζόταν ως $A^{'} : A_1 \times A_2 \times A_3 \times \cdots \times A_{|N|} = A^{|N|}$. Δεδομένου ότι οι έξοδοι του νευρωνικού για δεδομένη κατάσταση s αναπαριστούν τις συναρτήσεις $Q(s, a|\theta)$ ή $\pi(a/s, \theta)$ ανάλογα με τον

Σχήμα 4.3: Αρχικτεκτονική Πολυεπίπεδου Νευρωνικού Δικτύου Πρόσθιας Τροφοδότησης [Hay09].



αλγόριθμο εκπαίδευσης, το επίπεδο εξόδου θα έχει μέγεθος $|A|^{|N|}$. Για $|A| = 10$ και $|N| = 4$ το επίπεδο εξόδου έχει μέγεθος 10000. Η πιο συχνή αρχιτεκτονική νευρωνικών δικτύων θέλει τα κρυφά επίπεδα να έχουν αρκετές φορές περισσότερους νευρώνες από ότι το επίπεδο εξόδου. Αν υποθέσουμε ότι το τελευταίο κρυφό επίπεδο έχει διπλάσιο πλήθος νευρώνων από το επίπεδο εξόδου προκύπτουν $2 \cdot 10^{16}$ συναπτικά βάρη μόνο στην έξοδο του δικτύου. Είναι πρακτικά ανέφικτο να εκπαιδεύσουμε ένα τέτοιο νευρωνικό με του πόρους που διαθέτουμε. Επίσης όσο αυξάνεται το πλήθος των χρηστών στη κυψέλη, το μέγεθος του νευρωνικού αυξάνεται εκθετικά.

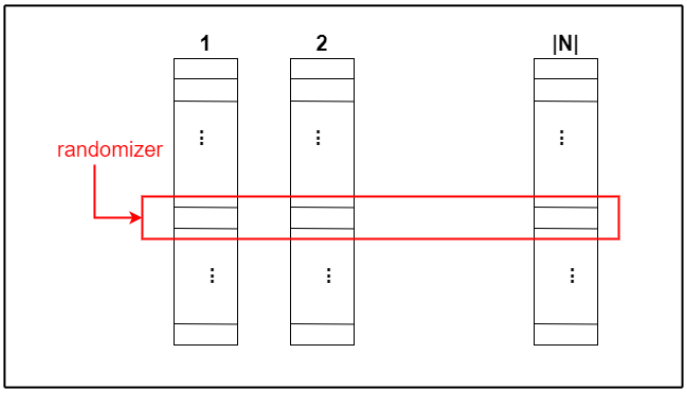
Στην περίπτωση των πολλαπλών πρακτόρων που ακολουθούμε στην παρούσα εργασία το επίπεδο εξόδου έχει μέγεθος $|A|$. Αυτό μας επιτρέπει να κατασκευάσουμε νευρωνικά μικρότερου μεγέθους και συνεπώς πιο εύκολα εκπαιδεύσιμα. Συγκεκριμένα, το νευρωνικό που προτείνουμε με 3 κρυφά επίπεδα έχει πλήθος παραμέτρων:

$$|\theta| = \sum_{i=0}^{3} (N_i + 1) N_{i+1}. \tag{4.17}$$

4.2.2 Deep Q-Learning (DQL)

Οι τεχνικές ενισχυτικής μάθησης που αναπαριστούν τη συνάρτηση Q με μη γραμμικές συναρτήσεις προσέγγισης, όπως τα νευρωνικά δίκτυα DQN (Deep Q Network), και χρησιμοποιούν μεθόδους Q-learning για την εύρεση της βέλτιστης Q^*, είναι ασταθείς. Οι αιτίες της αστάθειας είναι κυρίως τρεις: οι διαδοχικές καταστάσεις που παρατηρεί ο πράκτορας είναι

Σχήμα 4.4: Αρχιτεκτονική Experience Replay



υψηλά συσχετισμένες, μικρές αλλαγές στις παραμέτρους του νευρωνικού μπορεί να οδηγήσουν σε σημαντική αλλαγή στην πολιτική και οι συσχετίσεις μεταξύ των τιμών εκπαίδευσης Q_{train} και των τιμών στόχων Q_{target}. Στην εργασία [Mni+15], προτείνεται μια παραλαγή του Q-learning που χρησιμοποιεί δύο ιδέες κλειδιά. Στην πρώτη χρησιμοποιείται ένας μηχανισμός γνωστός και ως experience replay [Mni+15]. Ο μηχανισμός αυτός αποθηκεύει σε μια μονάδα μνήμης τις εμπειρίες που αποκτούν οι πράκτορες και σε κάθε βήμα εκπαίδευσης με τη βοήθεια ενός randomizer επιλέγει ένα minibatch εμπειριών D, ώστε να εξαλείψει τις συσχετίσεις που προκύπτουν όταν η εκπαίδευση γίνεται σε εμπειρίες που έχουν αποκτηθεί σε κοντινές χρονικές σχισμές. Η δεύτερη ιδέα προτείνει την κατασκευή δύο νευρωνικών δικτύων DQN_{train} και DQN_{target} τα οποία αναπαριστούν τις συναρτήσεις Q_{train} και Q_{target} αντίστοιχα. Η παράμετρος θ_{target} μένει σταθερή για ένα συγκεκριμένο διάστημα χρονικών σχισμών T_u και στο πέρας του διαστήματος αυτού ενημερώνεται ως $\theta_{target} = \theta_{train}$. Ο στόχος είναι να μειωθεί η εξάρτηση των τιμών Q_{train} και Q_{target}.

Για να εφαρμόσουμε experience replay χρησιμοποιούμε $|N|$ FIFO (First In-First Out) ουρές μεγέθους M για να αποθηκεύουμε τις εμπειρίες που αποκτούν οι $|N|$ πράκτορες. Σε κάθε πράκτορα αντιστοιχεί μία FIFO ουρά στην οποία ο πράκτορας i την χρονική σχισμή t αποθηκεύει την εμπειρία του $e_i^{(t)} = \left(s_i^{(t-1)}, a_i^{(t-1)}, r_i^{(t)}, s_i^{(t)}\right)$. Σε κάθε βήμα εκπαίδευσης, ένας κοινός randomizer επιλέγει $\frac{|D|}{|N|}$ εμπειρίες από κάθε ουρά και κατασκευάζει το minibatch $D^{(t)}$. Στο Σχήμα 4.4, απεικονίζεται η λειτουγία του experience replay. Ο randomizer παράγει ένα σύνολο δεικτών και στη συνέχεια επιλέγονται οι εμπειρίες που βρίσκονται στις αντίστοιχες θέσεις σε κάθε ουρά.

Όπως εξηγήθηκε και στην Ενότητα 2.3.4, για να εκπαιδευτούν οι παραμετροποιημένες συναρτήσεις τιμής χρειάζεται να ορίσουμε μία συνάρτηση σφάλματος πάνω σε κάποιο σύνολο δειγμάτων εκπαίδευσης. Ορίζουμε τη συνάρτηση σφάλματος για ένα minibatch ως:

$$L(\theta_{train}^t) = \sum_{(s,a,r',s') \in D^{(t)}} \left(y_{DQN} - Q(s,a|\theta_{train}^{(t)})\right)^2, \tag{4.18}$$

όπου

$$y_{DQN}^{(t)} = r' + \gamma \max_{a'} Q(s', a'|\theta_{target}^{(t)}). \tag{4.19}$$

Το $L(\theta^t_{train})$ ονομάζεται σφάλμα ελαχίστων τετραγώνων. Παρατηρούμε ότι ο ορισμός του y_{DQN} είναι ακριβώς ο ίδιος με τον ορισμό της τιμής στόχος Υ που είχε οριστεί για την απλή μέθοδο Q-learning που αναλύθηκε στην Ενότητα 2.3.3, δηλαδή η πολιτική στόχος (target policy) είναι να επιλεχθεί η βέλτιστη μέχρι εκείνη τη στιγμή ενέργεια.

Προκειμένου να παράξουμε μία καλύτερη εκτίμηση για την Q^*, σε κάθε βήμα εκπαίδευσης ενημερώνουμε την παράμετρο θ_{train} με τη μέθοδο της καθοδικής κλίσης (gradient descent):

$$\theta^{(t+1)}_{train} = \theta^{(t)}_{train} - \eta_q \nabla_{\theta_{train}} L(\theta^{(t)}_{train}). \tag{4.20}$$

Επίσης κατά τη διάρκεια της εκπαίδευσης υιοθετούμε μία ϵ-greedy πολιτική μάθησης (training/behavior policy) η οποία ελέγχει την πιθανότητα εξερεύνησης. Η παράμετρος ϵ περιγράφει την πιθανότητα να επιλεχθεί μία τυχαία ενέργεια αντί της βέλτιστης μέχρι εκείνη τη στιγμή ενέργειας και ορίζεται ως:

$$\epsilon_k = \epsilon_1 + \frac{k-1}{N_e - 1}(\epsilon_{N_e} - \epsilon_1) \quad k = 1, 2, ..., N_e. \tag{4.21}$$

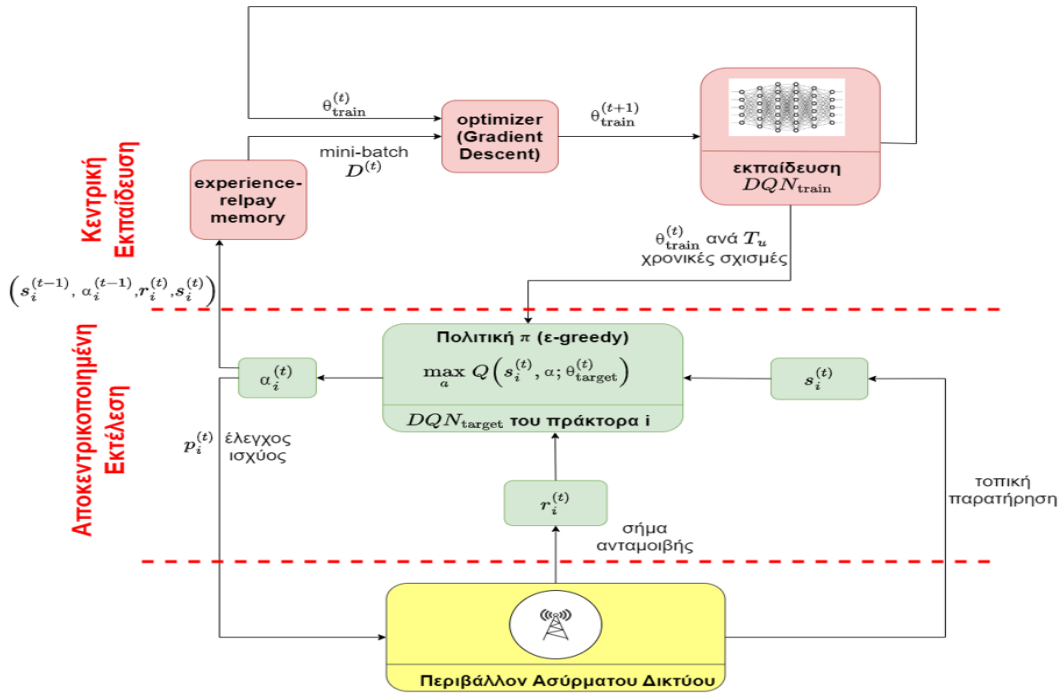
όπου N_e είναι το πλήθος των επεισοδίων, ϵ_1 η τιμή του ϵ στο πρώτο επεισόδιο και ϵ_{N_e} στο τελευταίο. Στα πρώτα επεισόδια η παράμετρος ϵ έχει μεγαλύτερη τιμή ώστε να εξερευνάται ο χώρος λύσεων, ενώ σταδιακά μειώνεται ώστε να εκμεταλλευτούμε τη γνώση που έχει αποκτηθεί.

Στο Σχήμα 4.5, απεικονίζεται γραφικά η φάση της εκπαίδευσης για τον πράκτορα i. Στην έναρξη της χρονικής σχισμής t, ο πράκτορας βρίσκεται στην τοπική κατάσταση $s^{(t)}_i$ και λαμβάνει ένα σήμα ανταμοιβής $r^{(t)}_i$. Επίσης, διαθέτει ένα αντίγραφο του DQN_{target} στο οποίο εκτελείται αποκεντρωμένα ο υπολογισμός των τιμών $Q_{target}(s^{(t)}_i, a|\theta_{target}), \forall a \in A$. Με βάση την πολιτική $\varepsilon-$greedy επιλέγεται η ενέργεια του πράκτορα. Στη συνέχεια, κατασκευάζεται η εμπειρία $e^{(t)}_i$ και αποθηκεύεται στην αντίστοιχη FIFO ουρά. Η διαδικασία αυτή συμβαίνει ταυτόχρονα για όλους του πράκτορες. Έπειτα, ο randomizer δημιουργεί ένα minibatch από εμπειρίες όλων των πρακτόρων. Τέλος, ενημερώνεται η παράμετρος θ_{train} με τη μέθοδο gradient descent. Την επόμενη χρονική στιγμή, ο πράκτορας παρατηρεί την τοπική κατάσταση $s^{(t+1)}_i$ και λαμβάνει σήμα ανταμοιβής $r^{(t+1)}_i$ και η διαδικασία επαναλαμβάνεται.

Στην περίπτωση όπου $\gamma = 0$, δηλαδή η επιλογή των ενεργειών γίνεται με σκοπό να μεγιστοποιήσουμε την άμεση ανταμοιβή, η τιμή στόχος y_{DQN} για μία εμπειρία $(s, a, r^{'}, s^{'})$ είναι $y_{DQN} = r^{'}$. Επομένως ο αλγόριθμος απλουστεύεται σημαντικά εφόσον δεν χρειάζεται να διατηρούμε αντίγραφο DQN_{target} ώστε να υπολογίσουμε το $\max_a Q(s^{'}, a|\theta_{target})$ ενώ οι εμπειρίες που αποθηκεύονται αρκεί να έχουν τη μορφή $(s, a, r^{'})$.

Διαισθητικά οι ενέργειες των πρακτόρων σε μία χρονική σχισμή οδηγούν σε μία λύση του προβλήματος βελτιστοποίησης του Κεφαλαίου 3 για τη χρονική σχισμή αυτή. Στην παρούσα εργασία, υποθέτουμε ότι δεν μας ενδιαφέρει η επίδραση που έχουν οι ενέργειες των πρακτόρων μίας δεδομένης χρονικής σχισμής, στις ενέργειες των επόμενων σχισμών εφόσον αυτές οδηγούν σε διαφορετικές λύσεις του προβλήματος βελτιστοποίησης. Επομένως, υλοποιούμε τον αλγόριθμο DQL υποθέτωντας $\gamma = 0$ και, συνεπώς, οι πράκτορες ενημερώνονται για το πόσο καλή είναι μία συγκεκριμένη λύση του προβλήματος βελτιστοποίησης αποκλειστικά από την άμεση ανταμοιβή. Η εφαρμογή του αλγορίθμου DQL περιγράφεται στον Αλγόριθμο 1.

Σχήμα 4.5: Απεικόνιση του προτεινόμενου πολλαπλών πρακτόρων DQL αλγορίθμου



4.2.3 REINFORCE

Η τεχνική REINFORCE είναι ένας policy search αλγόριθμος που αναπαριστά απευθείας τη στοχαστική πολιτική $\pi(a/s, \theta_\pi)$ με χρήση νευρωνικού δικτύου DPN (Deep Policy Network) (Ενότητα 2.3.4). Ο αλγόριθμος εκμάθησης βασίζεται σε εμπειρίες που λαμβάνονται με τη μέθοδο Monte-Carlo. Οι εμπειρίες αυτές αποκτώνται κατά τη διάρκεια κάθε επεισοδίου και είναι εμπειρικές, δηλαδή μέσοι όροι των δειγμάτων του επεισοδείου [Ben09]. Ο σκοπός τη χρονική σχιγμή t είναι να ενημερώσουμε την παράμετρο θ_π ώστε να μεγιστοποιήσουμε την αναμενόμενη μέση ανταμοιβή των πρακτόρων:

$$J(\theta_\pi^{(t)}) = \mathbb{E}_\pi \left[\frac{\sum_{i=1}^{|N|} r_i^{(t)}}{|N|} \right]. \tag{4.22}$$

Για να βρούμε το θ_π^* που μεγιστοποιεί το $J(\theta_\pi)$, χρησιμοποιούμε τη μέθοδο της ανοδικής κλίσης (Ενότητα 2.3.4) κατά την οποία:

$$\theta_\pi^{(t+1)} = \theta_\pi^{(t)} + \eta_\pi \nabla_{\theta_\pi} J(\theta_\pi^{(t)}). \tag{4.23}$$

όπου η_π ο ρυθμός μάθησης.

Με βάση το Policy Gradient Theorem [Ben09], προκύπτει:

$$\nabla_{\theta_\pi} J(\theta_\pi^{(t)}) = \frac{\sum_{i=1}^{|N|} \nabla_{\theta_\pi} \ln \pi(a_i^{(t)} | s_i^{(t)}, \theta_\pi^{(t)}) r_i^{(t)}}{|N|}. \tag{4.24}$$

Αλγόριθμος 1 Deep Q Learning

Είσοδος: Πλήθος επεισοδίων N_e, πλήθος χρονικών σχισμών επεισοδίου N_s, ρυθμός μάθησης η_q, αρχική και τελική πιθανότητα εξευρεύνησης ϵ_1 και ϵ_{N_e}, μέγεθος M ουρών FIFO, μέγεθος minibatch $|D|$.
Αρχικοποίηση: Αρχικοποίησε την παραμέτρο θ_q του DQN σε τυχαία τιμή.
for $k = 1$ to N_e **do**
  Ενημέρωσε το ϵ_k με βάση την Εξίσωση 4.21.
  Λάβε τις αρχικές καταστάσεις των πρακτόρων $s_i^0, \quad i = 1, 2, \ldots, |N|$.
  for $t = 1$ to N_s **do**
    for $i = 1$ to $|N|$ **do**
      if rand() $\leq \epsilon_k$ **then**
        Επίλεξε για τον πράκτορα i τυχαία μια ενέργεια $a_i^{(t)} \in A$.
      else
        Επίλεξε $a_i^{(t)} = arg\max_a Q(s_i^{(t)}, a/\theta_q^{(t)})$.
      end if
    end for
    Για $\mathbf{P}^{(t)} = [a_1^{(t)}, a_2^{(t)}, \ldots, a_{|N|}^{(t)}]$ υπολόγισε το $(\mathbf{c}^{(t)}, p_0^{(t)})$ που μεγιστοποιεί την EE
    Ανέθεσε στον BS τη λύση $(\mathbf{c}^{(t)}, \mathbf{P}^{(t)}, p_0^{(t)})$ και παρατήρησε τις νέες καταστάσεις $s_i^{(t+1)}, \quad i = 1, 2, \ldots, |N|$ και τις αμοιβές $r_i^{(t+1)}, \quad i = 1, 2, \ldots, |N|$.
    Κατασκεύασε τις εμπειρίες των πρακτόρων και τοποθέτησέ τες στις αντίστοιχες FIFO ουρές.
    Επίλεξε τυχαία ένα minibatch και υπολόγισε το $\nabla L(\theta_q^{(t)})$. Στη συνέχεια εκτέλεσε $\theta_q^{(t+1)} = \theta_q^{(t)} - \eta_q \nabla L(\theta_q^{(t)})$.
    Θέσε $s_i^{(t)} \leftarrow s_i^{(t+1)}, \quad i = 1, 2, \ldots, |N|$.
  end for
end for
Έξοδος: Εκπαιδευμένο DQN $Q(s, a|\theta_q)$

Εφόσον το νευρωνικό δίκτυο παράγει απευθείας τη στοχαστική πολιτική, οι ενέργειες που επιλέγουν οι πράκτορες τη χρονική σχισμή t ακολουθούν την πολιτική $\pi(a|s, \theta_\pi)$. Δηλαδή, ο πράκτορας i επιλέγει την ενέργεια $a_i^{(t)}$ με πιθανότητα $\pi(a_i^{(t)}|s_i^{(t)}, \theta_\pi^{(t)})$.

Ο αλγόριθμος είναι αρκετά ασταθής στη φάση της εκπαίδευσης, όταν οι ανταμοιβές έχουν μεγάλες αποκλίσεις μεταξύ τους. Προκειμένου να αντιμετωπίσουμε το πρόβλημα αυτό, πριν υπολογίσουμε το gradient, κανονικοποιούμε τις ανταμοιβές ως:

$$r_{i,norm}^{(t)} = \frac{r_i^{(t)} - \mu_r^{(t)}}{\sigma_r^t}, \tag{4.25}$$

όπου $\mu_r^{(t)} = \frac{\sum_{i=1}^{|N|} r_i^{(t)}}{|N|}$ είναι η μέση τιμή και $\sigma_r^{(t)} = \sqrt{\frac{\sum_{i=1}^{|N|}(r_i^{(t)} - \mu_r^{(t)})^2}{|N|}}$ η διασπορά των ανταμοιβών των πρακτόρων τη χρονική σχισμή t. Ο αλγόριθμος REINFORCE περιγράφεται στον Αλγόριθμο 2.

Αλγόριθμος 2 REINFORCE

Είσοδος: Πλήθος επεισοδίων N_e, πλήθος χρονικών σχισμών επεισοδίου N_s, ρυθμός μάθησης η_π.
Αρχικοποίηση: Αρχικοποίησε την παράμετρο θ_π του DPN σε τυχαία τιμή.
for $k = 1$ to N_e **do**
  Λάβε τις αρχικές καταστάσεις των πρακτόρων $s_i^0, \quad i = 1, 2, \ldots, |N|$.
  for $t = 1$ to N_s **do**
    Για κάθε πράκτορα i επίλεξε την ενέργεια που εκτελεί με βάση τις πιθανότητες $\pi(a/s_i^{(t)}, \theta_\pi^{(t)}), \quad i = 1, 2, \ldots, |N|$.
    Για $\mathbf{P}^{(t)} = [a_1^{(t)}, a_2^{(t)}, \ldots, a_{|N|}^{(t)}]$ υπολόγισε το $(\mathbf{c}^{(t)}, p_0^{(t)})$ που μεγιστοποιεί την WSR.
    Ανάθεσε στον BS τη λύση $(\mathbf{c}^{(t)}, \mathbf{P}^{(t)}, p_0^{(t)})$ και παρατήρησε τις νέες καταστάσεις $s_i^{(t+1)}, \quad i = 1, 2, \ldots, |N|$ και τις αμοιβές $r_i^{(t+1)}, \quad i = 1, 2, \ldots, |N|$.
    Υπολόγισε τα $\mu_r^{(t)}$ και $\sigma_r^{(t)}$ και στη συνέχεια τα $r_{i,norm}^{(t)}, \quad i = 1, 2, \ldots, |N|$.
    Υπολόγισε $\nabla J_{\theta_\pi^{(t)}}$ και ενημέρωσε $\theta_\pi^{(t+1)} = \theta_\pi^{(t+1)} + \eta_\pi \nabla \theta_{J_{\theta_\pi^{(t)}}}$.
    Θέσε $s_i^{(t)} \leftarrow s_i^{(t+1)}, \quad i = 1, 2, \ldots, |N|$
  end for
end for
Έξοδος: Εκπαιδευμένο DPN για $\pi(s|a, \theta_\pi)$

4.3 Μεγιστοποίηση Ρυθμού Μετάδοσης

4.3.1 Τροποποίηση μοντελοποίησης MDP

Στην ενότητα αυτή τροποποιούμε τη μοντελοποίηση της Ενότητας 4.1 ώστε να επιλύσουμε το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\max_{\boldsymbol{c},\boldsymbol{P},P_0} WSR = \frac{1}{|N|} \sum_{n=1}^{|N|} w_n R_n \tag{4.26}$$

$$\text{s.t.} \quad \sum_{n=1}^{|N|} c_n \leq r_1^c, \tag{4.27}$$

$$p_0 - \sum_{j=1}^{|N|} p_j \geq \frac{p_{tol} + \sigma^2}{g_1}, \tag{4.28}$$

$$\sum_{n=1}^{|N|} p_n + p_0 \leq P_{max}, \tag{4.29}$$

$$c_n, p_n \geq 0 \quad \forall n \quad \text{και} \quad p_0 \geq 0, \tag{4.30}$$

δηλαδή να μεγιστοποιήσουμε το WSR χωρίς να λάβουμε υπόψη το EE. Το WSR μετριέται σε bits/Hz.

Η τροποποίηση γίνεται ως εξής:

(i) Για κάθε ενέργεια $\boldsymbol{P}$ των πρακτόρων επιλέγουμε τη λύση $(\boldsymbol{c}[p_0], p_0 = p_{max} - \sum_{j=1}^{|N|} p_j)$. Ο λόγος είναι ότι το κοινό stream δεν δημιουργεί παρεμβολή στα ιδιωτικά streams με αποτέλεσμα η ανάθεση όλου του διαθέσιμου εύρους ισχύος στο κοινό stream να αυξάνει το WSR στην κυψέλη. Ο υπολογισμός του $\boldsymbol{c}[p_0]$ προκύπτει επιλύοντας το γραμμικό πρόβλημα βελτιστοποίησης 4.3.

(ii) Η συνάρτηση ανταμοιβής είναι ανάλογη του WSR και όχι του EE.

$$r_i^{(t)} = \frac{1}{|N|}\sum_{n=1}^{|N|} w_n R_n, \quad \text{εάν} \quad p_0 - \sum_{j=1}^{|N|} p_j \geq \frac{p_{tol}+\sigma^2}{g_1}, \tag{4.31}$$

$$r_i^{(t)} = \frac{1}{|N|}\sum_{n=1}^{|N|} w_n R_n(1+tanh(p_0-\sum_{j=1}^{|N|} p_j - \frac{p_{tol}+\sigma^2}{g_1})), \quad \text{εάν} \quad p_0-\sum_{j=1}^{|N|} p_j \leq \frac{p_{tol}+\sigma^2}{g_1}. \tag{4.32}$$

4.3.2 Αλγόριθμος WMSE

Ο αλγόριθμος αυτός είναι ένα κλασσικός αλγόριθμος ανάθεσης ισχύος για μεγιστοποίηση ρυθμού μετάδοσης. Δεν εφαρμόζει Rate-Splitting και στέλνει ολόκληρα τα μηνύματα μέσω ιδιωτικών ρυθμών. Ο αλγόριθμος αυτός χρησιμοποιείται ως σημείο αναφοράς (benchmark) για την αξιολόγηση της επίδοσης των αλγορίθμων ενισχυτικής μάθησης και περιγράφεται στον Αλγόριθμο 3.

Αλγόριθμος 3 Weighted Mean Square Error

1: Εισόδος: Μέγιστος σριθμός επαναλήψεων N_{max}
2: Αρχικοποίησε: v_i^0 για κάθε $i = 1, 2, ..|N|$ σε τυχαίες τιμές.
3: Υπολόγισε $u_i^0 = \frac{|h_i|v_i^0}{\sum_{j=1}|N||h_i|^2 v_j^0+\sigma^2}$.
4: Υπολόγισε $w_i^0 = \frac{1}{1-u_i^0|h_i|v_i^0}$.
5: Θέσε $C_{last} = \sum_{j=1}^{|N|} w_j^0$
6: **repeat**
7: $t = t + 1$
8: Υπολόγισε $v_i^{(t)} = \frac{a_i w_i^{(t-1)} u_i^{(t-1)}|h_i|}{\sum_{j=1}^{|N|}|h_i|^2 a_j w_j^{(t-1)}(u_j^{(t-1)})^2}$
9: Θέσε $v_i^{(t)} = \min(\sqrt{P_{max,i}}, v_i^{(t)})$.
10: Υπολόγισε $u_i^{(t)} = \frac{|h_i|v_i^{(t)}}{\sum_{j=1}|N||h_i|^2 v_j^{(t)}+\sigma^2}$.
11: Θέσε $C = \sum_{j=1}^{|N|} w_j^{(t)}$
12: Υπολόγισε Criterion $= C - C_{last}$.
13: Θέσε $C_{last} = C$.
14: **until** $Criterion \leq 10^{-3}$ ή $t = N_{max}$
15: Έξοδος: $p_i = v_i^2$ για κάθε $i = 1, 2, ..|N|$.

5

Αξιολόγηση Αλγορίθμων και Αριθμητικά Αποτελέσματα

5.1 Εισαγωγή

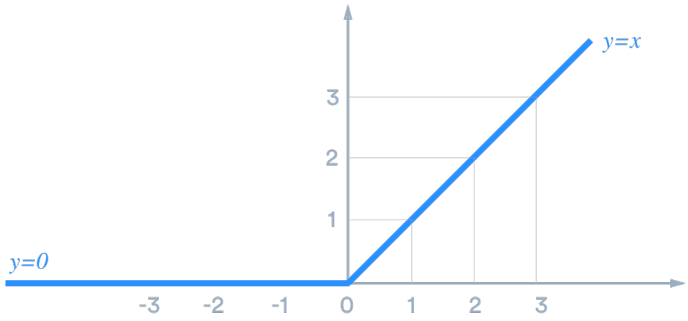
Στο κεφάλαιο αυτό θα αξιολογήσουμε μέσω προσομοιώσεων την απόδοση του μοντέλου διαχείρισης πόρων που προτείνουμε. Αρχικά, θα παρουσιαστεί η αξιολόγηση του αλγορίθμου DQL ως προς τη βελτιστοποίηση του EE και του WSR κατά τη φάση της εκπαίδευσης και στη συνέχεια σε δείγματα επικύρωσης (validation samples) ώστε να εξασφαλιστεί η ορθή λειτουργία του. Έπειτα, θα ακολουθήσει μια ξεχωριστή ανάλυση μεταξύ των δύο διαφορετικών στόχων βελτιστοποίησης. Προς βελτιστοποίηση του EE, ο αλγόριθμος DQL συγκρίνεται με τον REINFORCE και αξιολογείται η συμπεριφορά που έχουν σε δείγματα επικύρωσης ως προς τις μεταβολές κάποιων παραμέτρων του δικτύου που κρίθηκαν σημαντικές. Επίσης, αξιολογείται η επίδραση της βελτιστοποίησης του EE στο WSR του δικτύου. Προς βελτιστοποίηση του WSR, οι αλγόριθμοι DQL και REINFORCE συγκρίνονται με τον αλγόριθμο WMSE. Επίσης, αξιολογείται η επίδραση της βελτιστοποίησης του WSR στο EE του δικτύου.

5.2 Παράμετροι Προσομοίωσης

5.2.1 Περιβάλλον Ανάπτυξης

Για την προσομοίωση του συστήματος, χρησιμοποιήθηκε το Colab, το οποίο αποτελεί περιβάλλον για ανάπτυξη σημειωματάριων Jyputer (Jyputer Notebboks) και το οποίο εκτελείται στο cloud. Διατίθεται από τη Google και υποστηρίζει πολλές δημοφιλείς βιβλιοθήκες μηχανικής μάθησης που μπορούν εύκολα να φορτωθούν στα σημειωματάρια. Παρέχει επίσης τη δυνατότητα χρήσης GPU η οποία κρίνεται απαραίτητη στις περισσότερες εφαρμογές της μηχανικής μάθησης. Στις περισσότερες περιπτώσεις της δωρεάν έκδοσης του Colab χρησιμοποιούνται $K80$ GPUs και RAM των 12 GB. Κατά την ανάπτυξη του κώδικα της παρούσας εργασίας χρησιμοποιήθηκε κυρίως η βιβλιοθήκη Pytorch, καθώς και βοηθητικές συναρτήσεις από την Tensorflow.

Σχήμα 5.1: ReLU



5.2.2 Αρχιτεκτονική Νευρωνικού

Όπως εξηγήθηκε και στην Ενότητα 4.2.1, έχει επιλεχθεί ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης με 3 κρυφά επίπεδα. Το πλήθος των νευρώνων σε κάθε επίπεδο επιλέγεται ως $N_1 = 200$, $N_2 = 100$ και $N_3 = 40$. Επίσης, το επίπεδο εισόδου έχει $N_0 = 8$ νευρώνες, ένας νευρώνας για κάθε χαρακτηριστικό, ενώ για το επίπεδο εξόδου ισχύει $N_4 = |A|$. Επιλέγεται $|A| = 10$ στάθμες ισχύος για τα ιδιωτικά streams. Από την εξίσωση 4.17 προκύπτει ότι το νευρωνικό έχει $|\theta| = 26350$ συναπτικά βάρη. Όσον αφορά το κοινό stream θέτουμε $N_{p_0} = 100$.

Ως συνάρτηση ενεργοποίησης επιλέγεται η Rectified Linear Unit (ReLU) και έχει τη μορφή που φαίνεται στο Σχήμα 5.1.

5.2.3 Παράμετροι Περιβάλλοντος και Αλγορίθμων Επίλυσης

Για τις προσομοιώσεις που θα ακολουθήσουν, υποθέτουμε ότι οι χρήστες έχουν την ίδια επίδραση στον στόχο βελτιστοποίησης. Επομένως θεωρούμε $w_1 = w_2 = \cdots = w_n = 1$.

Στην Ενότητα 4.1.2, παραθέσαμε δύο διαφορετικές μεθόδους μοντελοποίησης της συνάρτησης ανταμοιβής. Μετά από αρκετές προσομοιώσεις αποδείχθηκε ότι η πρώτη μέθοδος δημιουργούσε αστάθεια κατά τη διάρκεια της εκπαίδευσης και τα αποτελέσματα ήτανε ακανόνιστα. Η πρώτη μέθοδος η οποία ήτανε και αρκετά πιο περίπλοκη δεν οδήγησε σε ορθή εκμαθηση των πρακτόρων. Ο λόγος που την παραθέτουμε είναι για να αναδείξουμε τη σημασία επιλογής της κατάλληλης συνάρτησης ανταμοιβής αλλά και να μεταφέρουμε τον σχετικό προβληματισμό στην ερευνητική κοινότητα. Οι προσομοιώσεις που παρατίθενται στην παρούσα εργασία υλοποιούν τη δεύτερη μέθοδο συνάρτησης ανταμοιβής.

Οι παράμετροι της προσομοίωσης του περιβάλλοντος που εκτελέστηκε για την εξαγωγή των αριθμητικών αποτελεσμάτων που θα ακολουθήσουν αρχικοποιήθηκαν στις τιμές που παρουσιάζονται στον πίνακα 5.1.

Στους πίνακες 5.2 και 5.3 παρουσιάζονται οι παράμετροι των αλγορίθμων επίλυσης του προβλήματος βελτιστοποίησης.

Πίνακας 5.1: Παράμετροι Προσομοίωσης Περιβάλλοντος

Παράμετρος	Τιμή
Πλήθος χρηστών	$\|N\| = 4$
Ελάχιστη απόσταση χρήστη από BS	$R_{min} = 10$ m
Μέγιστη απόσταση χρήστη από BS	$R_{max} = 1$ km
Διασπορά Doppler	$f_d = 10$ Hz
Διάρκεια Χρονικής Σχισμής	$T = 20$ ms
Additive White Gaussian Noise	$\sigma^2 = -114$ dBm
Απώλειες Σκίασης	$\sigma_2^2 = 8$ dB
Ευαισθησία δέκτη	$p_{tol} = -94$ dBm
Ελάχιστη ισχύς μετάδοσης μηνύματος	$P_{min} = 1$ dBm
Μέγιστη συνολική ισχύς μετάδοσης	$P_{max} = 40$ dBm

Πίνακας 5.2: Παράμετροι Αλγορίθμου DQL

Παράμετρος	Τιμή
Πλήθος επεισοδίων	$N_e = 1600$
Πλήθος χρονικών σχισμών ανά επεισόδιο	$N_s = 50$
Ρυθμός μάθησης	$\eta_q = 0.01$
Αρχική πιθανότητα εξερεύνησης	$\epsilon_1 = 0.2$
Τελική πιθανότητα εξερεύνησης	$\epsilon_{N_e} = 0.0001$
Μέγεθος FIFO ουρών για experience replay	$M = 5000$
Μέγεθος minibatch	$\|D\| = 500$

Πίνακας 5.3: Παράμετροι Αλγορίθμου REINFORCE

Παράμετρος	Τιμή
Πλήθος επεισοδίων	$N_e = 1600$
Πλήθος χρονικών σχισμών ανά επεισόδιο	$N_s = 50$
Ρυθμός μάθησης	$\eta_\pi = 0.01$

5.3 Αξιολόγηση Διαδικασίας Βελτιστοποίησης

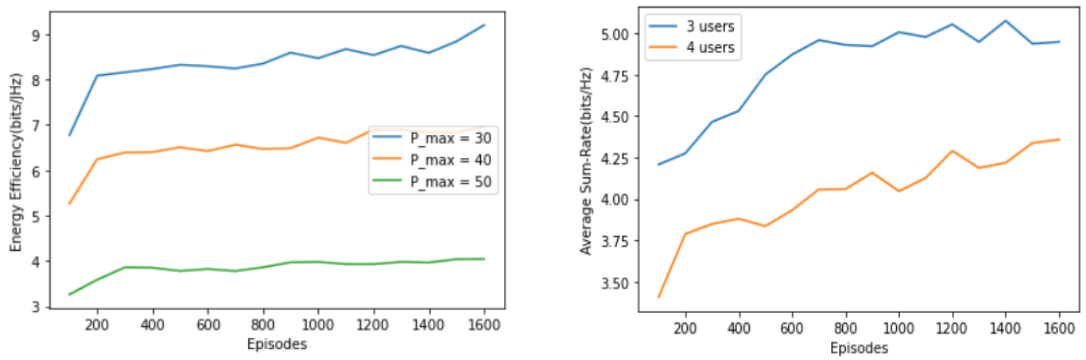
Σε αυτήν την ενότητα, μελετάται η απόδοση του αλγορίθμου DQL ως προς τους δύο διαφορετικούς στόχους βελτιστοποίησης του RMSA δικτύου. Σε πρώτη φάση αξιολογείται η συμπεριφορά του δικτύου ως προς τους στόχους βελτιστοποίησης κατά τη διάρκεια της εκπαίδευσης ενώ σε δεύτερη φάση αξιολογείται η επίδοση που έχουν τα εκπαιδευμένα πλέον νευρωνικά δίκτυα σε τυχαία δείγματα επικύρωσης.

5.3.1 Φάση Εκπαίδευσης

Στα παρακάτω γραφήματα παρουσιάζεται η μεταβολή του στόχου βελτιστοποίησης κατά τη διάρκεια της εκπαίδευσης για την τεχνική DQL.

Το EE (WSR) ενός επεισοδίου ορίζεται ως η μέση τιμή του EE (WSR) των N_s χρονικών σχισμών που συνθέτουν το επεισόδιο. Συγκεκριμένα το γράφημα 5.2α΄ απεικονίζει τη μεταβολή της ενεργειακής απόδοσης ως συνάρτηση του αριθμού των επεισοδίων. Οι διαφορετικές καμπύλες παρουσιάζουν μια ανάλυση ως προς τη μέγιστη εκπεμπόμενη ισχύ $P_{max} = [30, 40, 50]$. Όμοια, το γράφημα 5.2β΄ απεικονίζει τη μεταβολή της ρυθμαπόδοσης με ανάλυση ως προς το πλήθος των χρηστών στην κυψέλη $|N| = 3, 4$.

Παρατηρούμε ότι στα πρώτα επεισόδια ο στόχος βελτιστοποίησης αυξάνεται απότομα, ενώ με την πάροδο των επεισοδίων τείνει να σταθεροποιηθεί. Ο λόγος είναι ότι στα πρώτα επεισόδια έχουμε μεγάλη πιθανότητα εξερεύνησης, οπότε αρχικά οι ενημερώσεις της παραμέτρου θ_q δεν οδηγούν σε καλή αναπαράσταση της βέλτιστης $Q(s, a|\theta_q)$. Σταδιακά, όμως, το DQN μαθαίνει όλο και καλύτερες αναπαραστάσεις, ενώ σε συνδυασμό με την ελάττωση της πιθανότητας εξερεύνησης σε κάθε επεισόδιο έως την τελική τιμή ϵ_{N_e}, η παράμετρος θ_q τείνει να σταθεροποιηθεί σε μία τελική τιμή θ_q^*.



α΄: Βελτιστοποίηση: Ενεργειακή Απόδοση β΄: Βελτιστοποίηση: Ρυθμαπόδοση

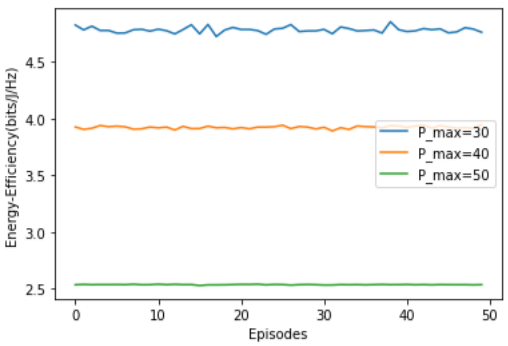
Σχήμα 5.2: Μεταβολή Στόχου Βελτιστοποίησης ως συνάρτηση του πλήθους των επεισοδίων κατά τη διάρκεια της εκπαίδευσης για την τεχνική DQL

5.3.2 Φάση Επικύρωσης

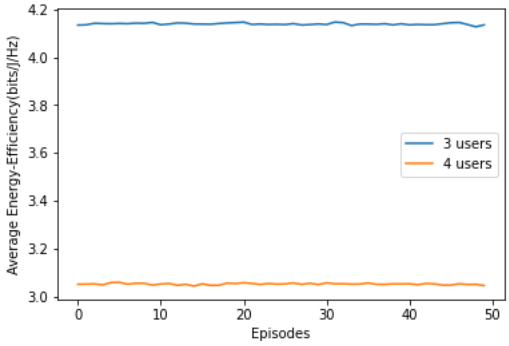
Στην υποενότητα αυτή, επικυρώνουμε εμπειρικά τη λειτουργικότητα του αλγορίθμου. Κατά τη διάρκεια της επικύρωσης, το κεντρικό δίκτυο εκπαίδευσης παύει να ενημερώνει τα συναπτικά βάρη του νευρωνικού ενώ ο αλγόριθμος $\epsilon-$greedy τερματίζεται, δηλαδή οι πράκτορες σταματάνε ολοκληρωτικά να εξερευνούν τον χώρο λύσεων. Η επικύρωση διαρκεί 50 επεισόδια και κάθε επεισόδιο διαρκεί 500 χρονικές σχισμές. Το EE (WSR) κάθε επεισοδίου προκύπτει ως η μέση τιμή του EE (WSR) των 500 χρονικών σχισμών που το συνθέτουν.

Στα γραφήματα 5.3α΄ και 5.3β΄ απεικονίζονται οι τιμές των στόχων βελτιστοποίησης για 50 διαφορετικά επεισόδια με ανάλυση ως προς τη μέγιστη ισχύ εκπομπής για το EE και το

πλήθος των χρηστών στη κυψέλη για το WSR. Παρατηρούμε ότι οι λύσεις θ_q που προέκυψαν στη φάση της εκπαίδευσης έχουν παρόμοιο EE και WSR για τυχαία δείγματα επικύρωσης, αποτέλεσμα που επιβεβαιώνει την ορθή λειτουργία της εκπαίδευσης.



α′: Βελτιστοποίηση: Ενεργειακή Απόδοση



β′: Βελτιστοποίηση: Ρυθμαπόδοση

Σχήμα 5.3: Μέση τιμή στόχου βελτιστοποίησης ανά επεισόδιο

5.4 Αξιολόγηση Διαφορετικών Στόχων Βελτιστοποίησης

Για να αποκτήσουμε περισσότερες πληροφορίες σχετικά με τη σημασία του στόχου βελτιστοποίησης της ενεργειακής απόδοσης για την συνολική επίδοση του ασυρμάτου δικτύου, προχωρούμε σε μια συγκριτική μελέτη μεταξύ των δύο ξεχωριστών στόχων βελτιστοποίησης. Συγκεκριμένα, εξετάζουμε τον τρόπο που μεταβάλλεται το EE και το WSR του δικτύου για διαφορετικές παραμέτρους του συστήματος.

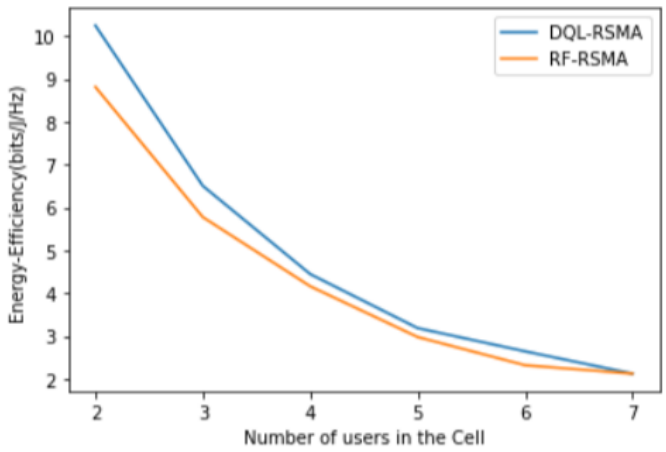
Η αξιολόγηση γίνεται με την ίδια ακριβώς διαδικασία που περιγράψαμε στην υποενότητα 5.3.2. Η αξιολόγηση γίνεται σε 50 ανεξάρτητα μεταξύ τους επεισόδια με 500 χρονικές σχισμές το κάθε ένα. Η τιμή του στόχου βελτιστοποίησης για κάθε επεισόδιο προκύπτει ως ο μέσος όρος της τιμής στόχου των χρονικών σχισμών που ανήκουν στο επεισόδιο αυτό. Η τελική τιμή του στόχου βελτιστοποίησης για την προσομοίωση προκύπτει ως ο μέσος όρος της τιμής στόχου των 50 επεισοδίων.

5.4.1 Βελτιστοποίηση Ενεργειακής Απόδοσης

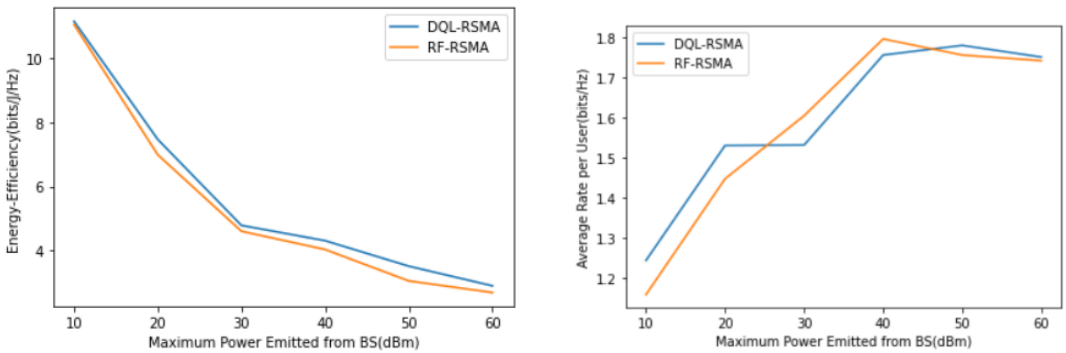
Στην υποενότητα αυτή εστιάζουμε αποκλειστικά στη βελτιστοποίηση της ενεργειακής απόδοσης και στην επίδραση που έχει στη ρυθμαπόδοση του δικτύου. Τα διαγράμματα που μελετάμε αφορούν τους αλγορίθμους DQL και REINFORCE.

Στο γράφημα 5.4 παρουσιάζεται η μείωση του EE του δικτύου καθώς οι χρήστες στο δίκτυο αυξάνονται. Παρατηρούμε ότι η προσθήκη ενός μόνο χρήστη έχει σημαντική επίδραση στο EE και συγκεκριμένα ο ρυθμός μεταβολής είναι μεγαλύτερος όταν οι χρήστες στο δίκτυο είναι λιγότεροι.

Σχήμα 5.4: Αξιολόγηση της ενεργειακής απόδοσης των αλγορίθμων DQL και REINFORCE για διαφορετικό πλήθος χρηστών στην κυψέλη



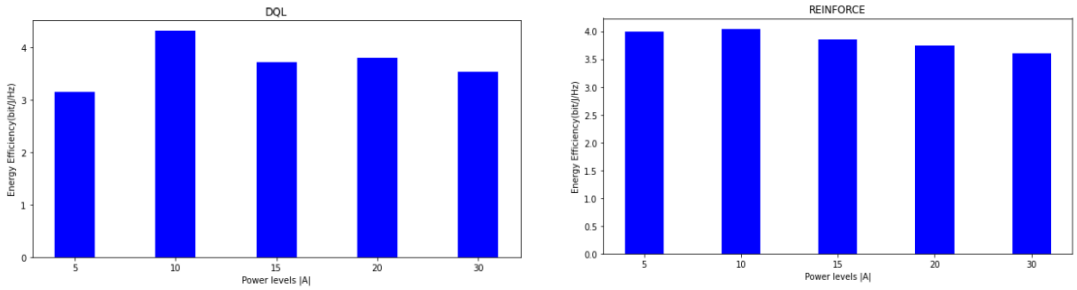
Στο γράφημα 5.5α′ απεικονίζεται το EE ως συνάρτηση της μέγιστης δυνατής εκπεμπόμενης ισχύος από τον BS, ενώ στο γράφημα 5.5β′ απεικονίζεται το αντίστοιχο WSR. Παρατηρούμε ότι παρόλο που η αύξηση της διαθέσιμης ισχύος οδηγεί σε λύσεις χαμηλότερης ενεργειακής απόδοσης οι ίδιες λύσεις οδηγούν σε αύξηση της ρυθμαπόδοσης. Ένα ενδιαφέρον αποτέλεσμα που προκύπτει από τη μελέτη των παραπάνω διαγραμμάτων είναι ότι από τα 40dBm και πάνω, το WSR εμφανίζει κορεσμό στο διάστημα $[1.7, 1.8]$bits/Hz, ενώ το EE του δικτύου συνεχίζει να μειώνεται με μικρότερο όμως ρυθμό μεταβολής. Επομένως, η εκπομπή με μέγιστη ισχύ άνω των 40dBm μειώνει την ενεργειακή απόδοση, χωρίς όμως αντίστοιχη συνεισφορά στη ρυθμαπόδοση.



α′: Ενεργειακή Απόδοση β′: Ρυθμαπόδοση

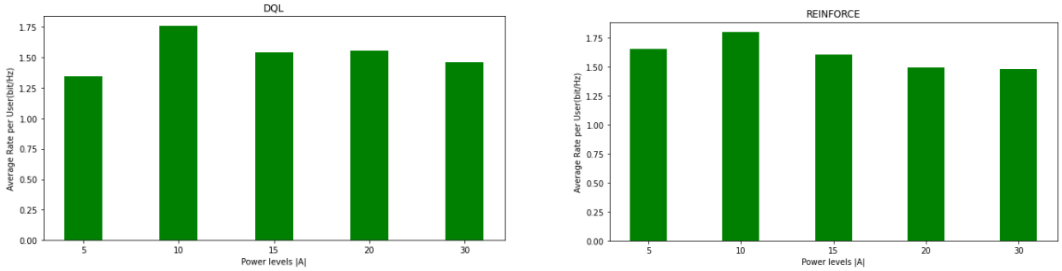
Σχήμα 5.5: Ενεργειακή απόδοση και ρυθμαπόδοση ως συνάρτηση της μέγιστης ισχύος εκπομπής

Στα Σχήματα 5.6 και 5.7 παρουσιάζεται η επίδραση που έχει το πλήθος των διακριτών επιπέδων ισχύος για τα ιδιωτικά μηνύματα στην αύξηση του στόχου βελτιστοποίησης. Ε-



α′: DQL β′: REINFORCE

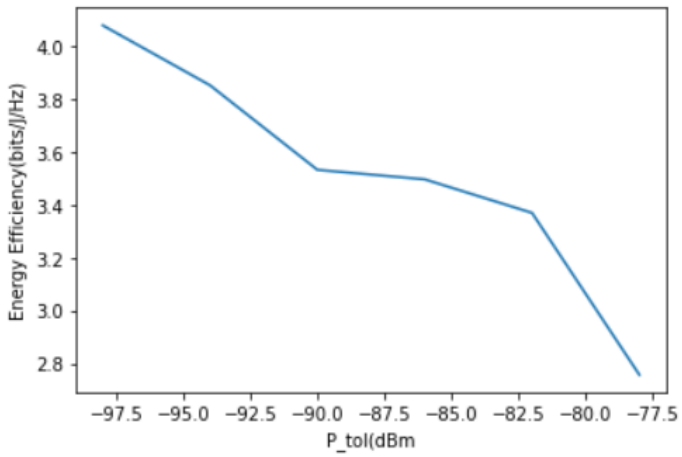
Σχήμα 5.6: Ενεργειακή απόδοση ως συνάρτηση του πλήθους των επιπέδων ισχύος ιδιωτικών μηνυμάτων



α′: DQL β′: REINFORCE

Σχήμα 5.7: Ρυθμαπόδοση ως συνάρτηση του πλήθους των επιπέδων ισχύος ιδιωτικών μηνυμάτων

Σχήμα 5.8: Εξάρτηση ενεργειακής απόδοσης από ευαισθησία δέκτη



ξάγουμε το συμπέρασμα ότι όταν το εύρος της ισχύος διαιρείται σε 10 στάθμες, οι λύσεις των αλγορίθμων DQL και REINFORCE έχουν υψηλότερη απόδοση από τις αντίστοιχες των ίδιων αλγορίθμων όταν εκτελούνται με λιγότερες ή περισσότερες στάθμες. Ο λόγος είναι ότι

για δεδομένη αρχιτεκτονική των κρυφών επιπέδων του προς εκπαίδευση νευρωνικού δικτύου, η αύξηση των νευρώνων εξόδου ενδέχεται να οδηγήσει σε αυξημένη πολυπλοκότητα, ενώ η μείωση αυτών μπορεί να οδηγήσει σε τετριμένες λύσεις.

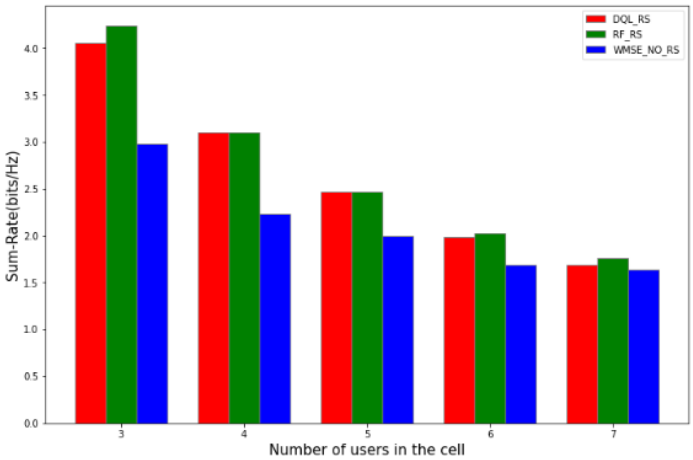
Στο γράφημα 5.8 απεικονίζεται η εξάρτηση της ενεργειακής απόδοσης από την ευαισθησία του δέκτη p_{tol} για τον αλγόριθμο DQL. Όπως έχουμε ήδη αναφέρει, το μέγεθος p_{tol} εκφράζει την ελάχιστη διαφορά ισχύος, στον δέκτη, μεταξύ κοινού stream και της συνολικής παρεμβολής από τα ιδιωτικά streams και τον θόρυβο καναλιού ώστε να μπορεί ο δέκτης να αποκωδικοποιήσει ορθά το κοινό μήνυμα. Παρατηρούμε ότι όσο απαιτούμε η διαφορά αυτή να είναι μεγαλύτερη, οι βέλτιστες λύσεις του αλγορίθμου DQL οδηγούν σε μειωμένη ενεργειακή απόδοση. Ο λόγος είναι ότι ο περιορισμός 3.16 σε συνδυασμό με μεγάλο p_{tol} καθιστά απαγορευτικές κάποιες λύσεις υψηλής ενεργειακής απόδοσης, οι οποίες σε περιπτώσεις χαμηλότερου p_{tol} θα ήτανε δεκτές.

5.4.2 Βελτιστοποίηση Ρυθμαπόδοσης

Στην υποενότητα αυτή εστιάζουμε στη βελτιστοποίηση της ρυθμαπόδοσης και στην επίδραση που έχει στην ενεργειακή απόδοση του δικτύου.

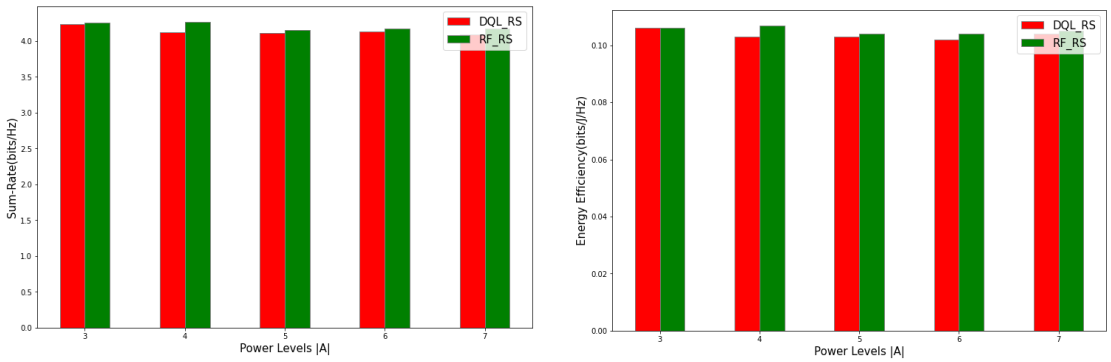
Επιπλέον των αλγορίθμων DQL και REINFORCE, προσομοιώνουμε και τον αλγόριθμο WMSE που περιγράψαμε στο 3. Για τον WMSE, θέτουμε $N_{max} = 100$, δηλαδή σε κάθε χρονική σχισμή ο αλγόριθμος εκτελείται 100 φορές ή και λιγότερες εάν ικανοποιείται το κριτήριο τερματισμού. Η έξοδος του αλγορίθμου είναι η ισχύς εκπομπής που ανατίθεται σε κάθε μήνυμα χρήστη κατά τη διάρκεια της επόμενης χρονικής σχισμής. Εκτελούμε τους αλγόριθμους για συνολικά 50 επεισόδια των 500 χρονικών σχισμών. Το μέσο WSR κάθε επεισοδίου ορίζεται ως ο μέσος όρος των ρυθμών μετάδοσης των 500 χρονικών σχισμών που συνθέτουν το επεισόδιο. Το τελικό WSR της προσομοίωσης ορίζεται ως ο μέσος όρος των 50 επεισοδίων.

Σχήμα 5.9: Αξιολόγηση Ρυθμαπόδοσης για διαφορετικό πλήθος χρηστών στο δίκτυο



Στο Σχήμα 5.9 απεικονίζεται η συνολική ρυθμαπόδοση ανά χρήστη για μεταβλητό πλήθος χρηστών στη κυψέλη. Παρατηρούμε ότι για λιγότερους χρήστες οι τεχνικές βαθιάς μάθησης

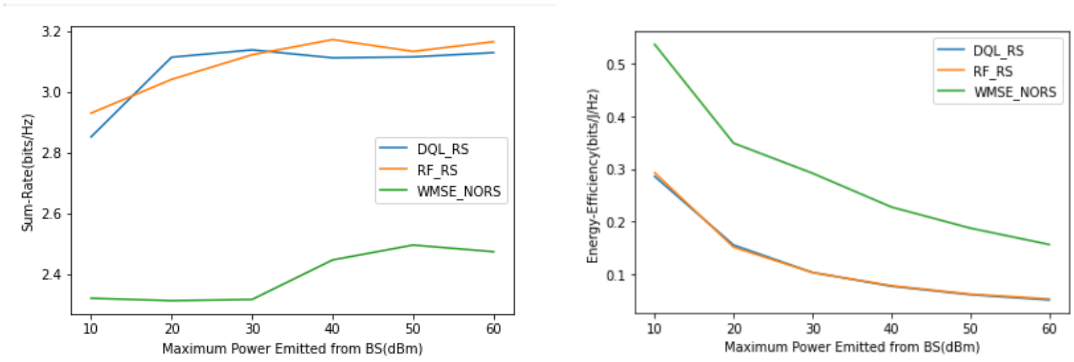
υπερέχουν σημαντικά του αλγορίθμου WMSE. Όσο προσθέτουμε όμως χρήστες στη κυψέλη ο WMSE αποδίδει περίπου όσο και οι τεχνικές βαθιάς μάθησης. Ο λόγος είναι ότι σε δίκτυο με περισσότερους χρήστες το πλήθος των πρακτόρων αυξάνεται με αποτέλεσμα η διαδικασία εκπαίδευσης να είναι πιο ασταθής αφού καλούμαστε να εκπαιδεύσουμε περισσότερους πράκτορες ταυτόχρονα. Επίσης παρατηρείται μείωση στο WSR όσο οι χρήστες στη κυψέλη αυξάνονται. Αυτό είναι αναμενόμενο αφού η προσθήκη χρήστη επιβαρύνει το δίκτυο.



α′: Ρυθμαπόδοση β′: Ενεργειακή Απόδοση

Σχήμα 5.10: Ρυθμαπόδοση και Ενεργειακή Απόδοση συναρτήσει του πλήθους των επιπέδων ισχύος των ιδιωτικών μηνυμάτων

Στο Σχήμα 5.10 περιγράφεται η εξάρτηση του ρυθμού μετάδοσης και της αντίστοιχης ενεργειακής απόδοσης από το πλήθος των επιπέδων ισχύος μετάδοσης των ιδιωτικών streams. Αντιλαμβανόμαστε από τα γραφήματα ότι οι διαφορές είναι αμελητέες, δηλαδή δεν παίζει κάποιο ρόλο η επιλογή του $|A|$ στη μεγιστοποίηση του WSR.

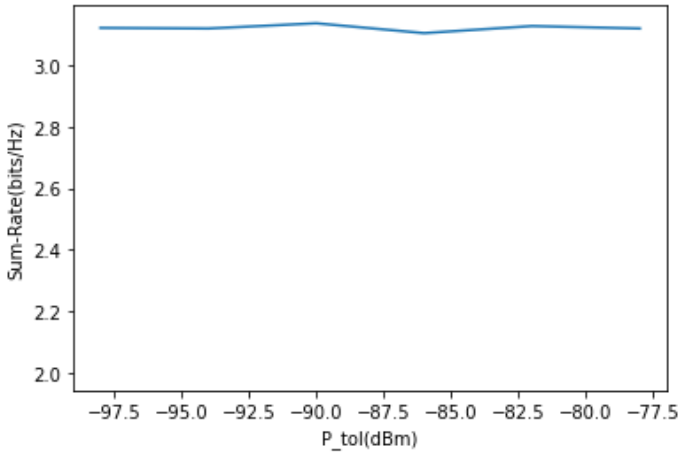


α′: Ρυθμαπόδοση β′: Ενεργειακή απόδοση

Σχήμα 5.11: Αξιολόγηση ρυθμού μετάδοσης και αντίστοιχης ενεργειακής απόδοσης συναρτήσει της μέγιστης διαθέσιμης ισχύος εκπομπής

Στο Σχήμα 5.11α′ απεικονίζεται ο τρόπος που μεταβάλλεται το WSR σε σχέση με τη μέγιστη διαθέσιμη ισχύ εκπομπής. Το Σχήμα 5.11β′ παρουσιάζει τις αντίστοιχες τιμές του EE. Παρατηρούμε ότι το WSR υφίσταται κορεσμό περί των 3.2bits/Hz για τις τεχνικές

Σχήμα 5.12: WSR εν συναρτήσει του p_{tol}



βαθιάς μάθησης ενώ για το WMSE ο κορεσμός συμβαίνει περί των 2.4 bits/Hz. Ταυτόχρονα παρατηρείται ελάττωση στην ενεργειακή απόδοση του δικτύου η οποία όμως σταθεροποιείται από τα 40dBm και πάνω. Συμπεραίνουμε λοιπόν ότι για $P_{max} \geq$ 40dBm οι λύσεις που προκύπτουν έχουν παρόμοιες αποδόσεις στο δίκτυο.

Όπως φαίνεται στο Σχήμα 5.12 ο ρυθμός μετάδοσης για την τεχνική DQL δεν εξαρτάται από την ευαισθησία του δέκτη. Αυτό συμβαίνει διότι προκειμένου να μεγιστοποιηθεί ο ρυθμός μετάδοσης το κοινό stream μεταδίδεται με σχετικά μεγάλη ισχύ ενώ τα ιδιωτικά με μικρότερη. Επομένως η παρεμβολή στο κοινό stream είναι σχετικά μικρή και η αποκωδικοποίησή του στους δέκτες των χρηστών είναι εφικτή ακόμα και σε δέκτες με μικρή ευαισθησία.

6

Σύνοψη - Συμπεράσματα

6.1 Σύνοψη

Μεγάλο πλήθος τεχνολογιών εμφανίστηκαν πρόσφατα, οι οποίες διαμορφώνουν σταδιακά την εποχή των μελλοντικών ασύρματων δικτύων. Πιθανοί υποψήφιοι τέτοιων τεχνολογιών αποτελούν εναλλακτικά σχήματα πολλαπλής πρόσβασης που μπορούν να προσφέρουν μεγαλύτερη χωρητικότητα στο δίκτυο. Στην παρούσα διπλωματική, προσδιορίσαμε τις δυνατότητες που προσφέρει το σχήμα Πολλαπλής Πρόσβασης Διαίρεσης Ρυθμού (RSMA) σε ένα δίκτυο μονής κυψέλης (single-cell) και Μίας Εισόδου-Μίας Εξόδου (SISO). Στόχος ήταν η βελτίωση της συνολικής ενεργειακής απόδοσης του δικτύου. Για το σκοπό αυτό, αναπτύχθηκε ένα δυναμικό μοντέλο διαχείρισης πόρων, που βασίστηκε κυρίως στις έννοιες της παρατήρησης και της μάθησης, το οποίο ελέγχει τις τιμές της ισχύος μετάδοσης του κοινού και των ιδιωτικών streams. Για να αντιμετωπίσουμε το πρόβλημα της ενεργειακής αποδοτικότητας στο δίκτυο, προτείναμε μία διαδικασία βελτιστοποίησης η οποία ακολουθεί τις αρχές της Ενισχυτικής Μάθησης και πραγματοποιεί την κατανομή των πόρων σε πραγματικό χρόνο.

6.2 Συμπεράσματα

Από την υλοποίηση των αλγοριθμικών μοντέλων που περιγράφηκαν και την πραγματοποίηση μιας σειράς προσομοιώσεων προέκυψε ένα μεγάλο πλήθος αριθμητικών αποτελεσμάτων που οδηγεί στα συμπεράσματα που ακολουθούν:

1. Το αλγοριθμικό μοντέλο για το DQL που παρουσιάστηκε συγκλίνει σε μία σταθερή λύση $\theta_\pi \rightarrow \theta_\pi^*$.

2. Η γενίκευση σε δείγματα επικύρωσης είναι αποδεκτή εφόσον επιτυγχάνεται παρόμοια τιμή του στόχου βελτιστοποίησης για διαφορετικά επεισόδια επικύρωσης. Αυτό σημαίνει ότι το νευρωνικό δίκτυο είναι καλά εκπαιδευμένο ώστε ανεξάρτητα των συνθηκών ενός δικτύου πραγματικού χρόνου, η ανάθεση των πόρων να προσαρμόζεται με τέτοιο τρόπο ώστε να μεγιστοποιείται το EE ή το WSR.

3. Από σύγκριση με την εργασία [MCL18] συμπεραίνουμε ότι τα αποτελέσματα μας για το ΕΕ βρίσκονται εντός ενός αποδεκτού εύρους τιμών. Συγκεκριμένα, η [MCL18] βρίσκει ΕΕ στο εύρος $(0-6)$bits/J/Hz για διάφορες προσομοιώσεις, ενώ τα αποτελέσματα της παρούσας εργασίας ως προς το ΕΕ κυμαίνονται στο διάστημα $(0-10)$bits/J/Hz για τα διάφορα πειράματα που εκτελέσαμε. Αν και τα αποτελέσματα της παρούσας εργασία δεν είναι απευθείας συγκρίσιμα με αυτά άλλων εργασιών, λόγω του ότι τα προς μελέτη τηλεπικοινωνιακά συστήματα παρουσιάζουν σημαντικές διαφορές (π.χ. στην παρούσα εργασία θεωρούμε σύστημα πραγματικού χρόνου), σχετικές εργασίες επιβεβαιώνουν ότι τα αποτελέσματα μας διαμορφώνονται σε αντίστοιχα επίπεδα. Ενδεικτικά αναφέρουμε ότι στην εργασία [Hie+21] το βέλτιστο WSR κυμαίνεται στο εύρος $(0-5.5)$bits/Hz. Στην παρούσα εργασία το εύρος βέλτιστου WSR είναι $(0-4.5)$bits/Hz.

4. Η αύξηση των πρακτόρων στο αλγοριθμικό μοντέλο βελτιστοποίησης μειώνει σημαντικά την επίδοση. Όταν έχουμε 3 ή 4 πράκτορες, οι τεχνικές ενισχυτικής μάθησης υπερισχύουν αρκετά του κλασσικού αλγορίθμου WMSE, ενώ όταν οι πράκτορες γίνονται 7 τότε οι επιδόσεις των τριών αλγορίθμων είναι παρόμοιες.

5. Η αύξηση του πλήθους των χρηστών επιβαρύνει σημαντικά το δίκτυο ως προς τα ΕΕ και WSR.

6. Οι αλγόριθμοι DQL και REINFORCE έχουν αρκετά παρόμοια συμπεριφορά. Παρόλα αυτά, παρατηρήθηκε ότι ο DQL αποδίδει ελαφρώς καλύτερα όταν επιθυμούμε να μεγιστοποιήσουμε το ΕΕ, ενώ ο REINFORCE όταν επιθυμούμε να μεγιστοποιήσουμε το WSR.

7. Ο αλγόριθμος WMSE έχει χειρότερη επίδοση από τις τεχνικές βαθιάς μάθησης ως προς το WSR αλλά πετυχαίνει καλύτερο ΕΕ όταν ο στόχος βελτιστοποίησης είναι το WSR. Παρατηρούμε ένα trade-off μεταξύ ΕΕ και WSR, δηλαδή οι αλγόριθμοι που πετυχαίνουν καλύτερο WSR έχουν χειρότερη επίδοση ως προς το ΕΕ. Αυτό συμβαίνει γιατί συνήθως οι λύσεις που δίνουν μεγαλύτερο WSR έχουν μεγαλύτερη κατανάλωση ισχύος. Επομένως όσο αναζητούμε λύσεις μεγαλύτερου WSR είναι αρκετά πιθανόν να οδηγήσουν σε χειροτέρευση του ΕΕ.

8. Η ευαισθησία του δέκτη p_{tol} επηρεάζει σε μεγάλο βαθμό το ΕΕ όταν ο στόχος βελτιστοποίησης είναι η μεγιστοποίηση του ίδιου του ΕΕ. Όταν όμως η βελτιστοποίηση αφορά τη μεγιστοποίηση του WSR, το p_tol δεν επηρεάζει το WSR του δικτύου.

9. Η μέγιστη διαθέσιμη ισχύς που εκπέμπει ο σταθμός βάσης παίζει καθοριστικό ρόλο στη διαμόρφωση της ΕΕ του δικτύου. Συγκεριμένα η αύξηση της μέγιστης διαθέσιμης ισχύος εκπομπής οδηγεί σε λύσεις μικρότερης ενεργειακής απόδοσης αλλά μεγαλύτερης ρυθμαπόδοσης. Παρατηρείται, όμως, ένα σημείο κορεσμού στη ρυθμαπόδοση, δηλαδή υπάρχει ένα μέγιστο όριο στη ρυθμαπόδοση που μπορεί να επιτύχει ένα δίκτυο τη στιγμή που στοχεύει σε μεγιστοποίηση του ΕΕ. Το ίδιο φαινόμενο παρατηρείται όταν το δίκτυο

προσπαθεί να μεγιστοποιήσει το WSR, δηλαδή η μεγιστοποίηση του WSR έχει αρνητική επίδραση στο EE.

10. Όταν ο στόχος είναι η μεγιστοποίηση του WSR το δίκτυο μπορεί να επιτύχει έως και 3 φορές μεγαλύτερη ρυθμαπόδοση σε σχέση με την ρυθμαπόδοση που παρατηρείται στο δίκτυο όταν μας ενδιαφέρει η μεγιστοποίηση του EE. Η αντίστοιχη, όμως, ενεργειακή απόδοση μειώνεται έως και 20 φορές. Είναι αρκετά δυσανάλογη η μείωση της EE ώστε να επιτευχθεί μεγαλύτερη ρυθμαπόδοση στο δίκτυο.

11. Η επιλογή της βελτιστοποίησης της ενεργειακής απόδοσης του δικτύου προσφέρει συνολικά περισσότερα οφέλη σε σχέση με την βελτιστοποίηση της ρυθμαπόδοσης στην κυψέλη.

6.3 Μελλοντική Έρευνα

Μελλοντικές επεκτάσεις της παρούσας εργασίας περιλαμβάνουν την επέκταση του προβλήματος σε δίκτυα MISO, όπου ο πομπός διαθέτει πολλαπλές κεραίες εκπομπής. Σε τέτοια δίκτυα το πρόβλημα βελτιστοποίησης αποκτά μεγαλύτερη πολυπλοκότητα γιατί ο αλγόριθμος επίλυσης αναζητά τα βέλτιστα beamforming vectors που αντιστοιχούν στις προς μετάδοση ροές δεδομένων. Επίσης, προτείνεται η επέκταση σε δίκτυα πολλαπλών κυψελών ώστε να εξεταστεί η επίδοση των τεχνικών που χρησιμοποιήθηκαν στην παρούσα εργασία σε περιπτώσεις που λαμβάνεται υπόψιν η διακυψελική παρεμβολή.

Όσον αφορά το αλγοριθμικό μοντέλο, μελλοντικές έρευνες θα μπορούσαν να εξετάσουν αλγορίθμους ενισχυτικής μάθησης με εφαρμογή σε συνεχείς χώρους λύσεων. Επίσης, μελλοντικές έρευνες θα μπορούσαν να αφορούν την αναζήτηση νέων τεχνικών μηχανικής μάθησης που να ενσωματώνουν κάτω από ένα κοινό πλαίσιο περισσοτέρων του ενός στόχων βελτιστοποίησης. Μπορεί, επίσης, να εξεταστούν νέες μοντελοποιήσεις του MDP της Ενότητας 4.1. Συγκεκριμένα, προτείνεται να εξεταστούν διάφορες μοντελοποιήσεις των καταστάσεων των πρακτόρων και της συνάρτησης ανταμοιβής. Επιπλέον, μπορούν να εξεταστούν νέες τεχνικές συστημάτων πολλαπλών πρακτόρων οι οποίες να ενσωματώνουν το κοινό stream σε ένα επιπρόσθετο πράκτορα.

Βιβλιογραφία

[Ahm+19] A. Alameer Ahmad, H. Dahrouj, A. Chaaban, A. Sezgin, and M. Alouini. "Interference mitigation via rate-splitting and common message decoding in cloud radio access networks". In: *IEEE Access* 7 (2019), pp. 80 350–80 365.

[Ahm+20] A. A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx. ""Rate splitting multiple access in C-RAN". In: *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)* (2020).

[al15] V. Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (2015), p. 529.

[al17] D. Silver et al. "Mastering the game of go without human knowledge". In: *Nature* 550.7676 (2017), pp. 354–359.

[ALH21] A. Alkhatee, G. Leus, and R. W. Heath. "Limited feedback hybrid precoding for multi-user millimeter wave systems". In: *IEEE Trans. Wireless Commun.* 14.11 (2021), pp. 6481–6494.

[AMK21] M. R. Camana Acosta, C. E. G. Moreta, and I. Koo. "Joint power allocation and power splitting for MISO-RSMA cognitive radio systems with SWIPT and information decoder users". In: *IEEE Syst. J.* 15.4 (2021), pp. 5289–5300.

[ATH16] M. S. Ali, H. Tabassum, and E. Hossain. "Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems". In: *IEEE Access* 4 (2016), pp. 6325–6343.

[Ben09] Y. Bengio. "Learning deep architectures for AI. Foundations and Trends in Machine Learning 2". In: (2009), pp. 1–127.

[BN18] M. Bennis and D. Niyato. "A Q-learning based approach to interference avoidance in self-organized femtocell networks". In: *Proc. IEEE Globecom Workshops* (2018), pp. 706–710.

[Buz+16] S. Buzzi, C. L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone. "A survey of energy-efficient techniques for 5G networks and challenges ahead". In: *IEEE Journal on Selected Areas in Communications* 34.4 (2016), pp. 697–709.

[CC15] Y. Wu C. Hao and B. Clerckx. "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach". In: *IEEE Trans. Commun.* 63.9 (Sept. 2015), pp. 3232–3246.

[Che+21] H. Chen, D. Mi, T. Wang, Z. Chu, Y. Xu, D. He, and P. Xiao. "Ratesplitting for multicarrier multigroup multicast: Precoder design and error performance". In: *IEEE Trans. Broadcast.* 67.3 (2021), pp. 619–630.

[Cle+19] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor. "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis". In: *arXiv preprint arXiv:1906.04474* (2019). URL: `http://arxiv.org/abs/1906.04474`.

[Cuc+11] Giuseppe Cuccu, Matthew Luciw, Jurgen Schmidhuber, and Faustino Gomez. "Intrinsically Motivated Neuroevolution for Vision-Based Reinforcement Learning". In: *ICDL* 2 (2011).

[Dai+15] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang. "Nonorthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends". In: *IEEE Commun. Mag.* 53.9 (Sept. 2015), pp. 74–81.

[DC17] M. Dai and B. Clerckx. "Multiuser millimeter wave beamforming strategies with quantized and statistical CSIT". In: *IEEE Trans. Wireless Commun.* 16.11 (2017), pp. 7025–7038.

[DC21] O. Dizdar and B. Clerckx. "Rate splitting multiple access for multiantenna multi-carrier joint communications and jamming". In: *Int. Conf. in Sens. Signal Process. for Defence (SSPD)* (2021).

[DHM06] N. Dawod, R. Hafez, and I. Marsland. "A multiuser zeroforcing system with reduced near-far problem and MIMO channel correlations". In: *Canadian Conf. on Electrical and Computer Engg.* (2006), pp. 936–939.

[DMC21] O. Dizdar, Y. Mao, and B. Clerckx. "Rate-splitting multiple access to mitigate the curse of mobility in (massive) MIMO networks". In: *IEEE Trans. Commun.* 69.10 (2021), pp. 6765–6780.

[Foe+18] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. "Counterfactual multiagent policy gradients". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (2018), pp. 2974–2982.

[GEK17] J. K. Gupta, M. Egorov, and M. Kochenderfer. "Cooperative multi-agent control using deep reinforcement learning". In: *International Conference on Autonomous Agents and Multiagent Systems* (2017), pp. 66–83.

[Has+21α] M. Z. Hassan, M. J. Hossain, J. Cheng, and V. C. M. Leung. "Device-clustering and rate-splitting enabled device-to-device cooperation framework in fog radio access network". In: *IEEE Trans. Green Commun. Netw.* 5.3 (2021), pp. 1482–1501.

[Has+21β] M. Z. Hassan, M. J. Hossain, J. Cheng, and V. C. M. Leung. "Energyspectrum efficient content distribution in Fog-RAN using rate-splitting, common message decoding, and 3d-resource matching". In: *IEEE Trans. Wireless Commun.* 20.8 (2021), pp. 4929–4946.

[Hay09] Simon Haykin. "Neural Networks and Learning Machines Third Edition". In: (2009).

[Hie+21] N. Q. Hieu, D. T. Hoang, D. Niyato, and D. I. Kim. "Optimal Power Allocation for Rate Splitting Communications With Deep Reinforcement Learning". In: *IEEE Wireless Communications Letters* 10.12 (2021), pp. 2820–2823.

[HW98] J. Hu and M. P. Wellman. "Online learning about other agents in a dynamic multiagent system". In: *International Conference on Autonomous Agents: Proceedings of the second international conference on Autonomous agents* 10.13 (1998), pp. 239–246.

[Jaa+20α] W. Jaafar, S. Naser, S. Muhaidat, P. C. Sofotasios, and H. Yanikomeroglu. "Multiple access in aerial networks: From orthogonal and non-orthogonal to rate-splitting". In: *IEEE Open J. of Veh. Technol.* 1 (2020), pp. 372–392.

[Jaa+20β] W. Jaafar, S. Naser, S. Muhaidat, P. C. Sofotasios, and H. Yanikomeroglu. "On the downlink performance of RSMA-based UAV communications". In: *IEEE Trans. Veh. Technol.* 69.12 (2020), pp. 16 258–16 263.

[JCS20] O. Dizdar J. An, B. Clerckx, and W. Shin. ""Rate-splitting multiple access for multi-antenna broadcast channel with imperfect CSIT and CSIR". In: *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)* (2020).

[JH07] J. Jiang M.and Akhtman and L. Hanzo. "Iterative joint channel estimation and multi-user detection for multiple-antenna aided OFDM systems". In: *EEE Trans. Wireless Commun.* 6(8) (2007), pp. 2904–2914.

[KB16] L. Kraemer and B. Banerjee. "Multi-agent reinforcement learning as a rehearsal for decentralized planning". In: *Neurocomputing* 190 (2016), pp. 82–94.

[Kou+13] Jan Koutnik, Giuseppe Cuccu, Jurgen Schmidhuber, and Faustino Gomez. "Evolving Large-Scale Neural Networks for Vision-Based Reinforcement Learning". In: *GECCO* (2013).

[Lan+17] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, J. Perolat, D. Silver, and T. Graepel. "A unified game-theoretic approach to multiagent reinforcement learning". In: *Advances in Neural Information Processing Systems* (2017), pp. 4193–4206.

[Lar+14] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta. "Massive MIMO for next generation wireless systems". In: *IEEE Communications Mag.* 52(2) (2014), pp. 186–195.

[Li+18] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li. "Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach". In: *IEEE Access* 6 (2018), pp. 25463–25473.

[Li+20] Z. Li, C. Ye, Y. Cui, S. Yang, and S. Shamai. "Rate splitting for multiantenna downlink: Precoder design and practical implementation". In: *IEEE J. Sel. Areas Commun.* 38.8 (2020), pp. 1910–1924.

[Li17] Y. Li. "Deep reinforcement learning: An overview". In: *CoRR* abs/1701.07274 (2017), pp. 1–85.

[Lia+18] F. Liang, C. Shen, W. Yu, and F. Wu. "Towards optimal power control via ensembling deep neural networks". In: *CoRR* abs/1807.10025 (2018), pp. 1–30.

[Lin+21] Z. Lin, M. Lin, B. Champagne, W.-P. Zhu, and N. Al-Dhahir. "Secure and energy efficient transmission for RSMA-based cognitive satelliteterrestrial networks". In: *IEEE Wireless Communications Letters* 10.2 (2021), pp. 251–255.

[LL08] G. Levin and S. Loyka. "On the Outage Capacity Distribution of Correlated Keyhole MIMO Channels". In: *IEEE Trans. Info. Theory* 54(7) (2008), pp. 3232–3245.

[LT02] A. Lozano and A. Tulino. "Capacity of multipletransmit multiple-receive antenna architectures". In: *IEEE Transactions on Information Theory* 48(12) (2002), pp. 3117–3128.

[Mao+20] Y. Mao, B. Clerckx, J. Zhang, V. O. K. Li, and M. A. Arafah. "Maxmin fairness of K-user cooperative rate-splitting in MISO broadcast channel with user relaying". In: *IEEE Trans. Wireless Commun.* 19.10 (2020), pp. 6362–6376.

[Mao+22] Yijie Maoa, Onur Dizdar, Bruno Clerckx, Robert Schober, Petar Popovski, and H. Vincent Poor. "Rate-Splitting Multiple Access: Fundamentals, Survey, and Future Research Trends". In: *arXiv:2201.03192 [cs.IT]* (2022).

[MC20α] Y. Mao and B. Clerckx. "Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: A rate-splitting approach". In: *IEEE Trans. Commun.* 68.11 (2020), pp. 6775–6791.

[MC20β] Y. Mao and B. Clerckx. "Dirty paper coded rate-splitting for nonorthogonal unicast and multicast transmission with partial CSIT". In: *Proc. 54th Asilomar Conf. Signals, Syst. Comput.* (2020).

[MCL18] Y. Mao, B. Clerckx, and V. O. Li. "Energy efficiency of rate-splitting multiple access, and performance benefits over sdma and noma". In: *15th ISWCS* (2018), pp. 1–5.

[MCW19] F. Meng, P. Chen, and L. Wu. "Power Allocation in Multi-User Cellular Networks with Deep Q Learning Approach". In: *IEEE International Conference on Communications (ICC)* (2019), pp. 1–6.

[Men+20] F. Meng, P. Chen, L. Wu, and J. Cheng. "Power Allocation in Multi-User Cellular Networks: Deep Reinforcement Learning Approaches". In: *EEE Transactions on Wireless Communications* 19.10 (2020), pp. 6255–6267.

[Mis+22α] A. Mishra, Y. Mao, L. Sanguinetti, and B. Clerckx. "Rate-splitting assisted massive machine-type communications in cell-free massive MIMO". In: *IEEE Commun. Lett.* 26.6 (2022), pp. 1358–1362.

[Mis+22β] A. Mishra, Y. Mao, C. K. Thomas, L. Sanguinetti, and B. Clerckx. "Mitigating intra-cell pilot contamination in massive MIMO: A rate splitting approach". In: *arXiv preprint arXiv:2206.07499* (2022).

[MK10] T. Maciel and A Klein. "On the performance, complexity, and fairness of suboptimal resource allocation for multiuser MIMO-OFDMA systems". In: *IEEE Trans. on Veh. Technol.* 59(1) (2010), pp. 406–419.

[Mni+15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (2015), pp. 529–533.

[MPC21] Y. Mao, E. Piovano, and B. Clerckx. "Rate-splitting multiple access for overloaded cellular internet of things". In: *IEEE Trans. Commun.* 69.7 (2021), pp. 4504–4519.

[NB13] H. Nikopour and H. Baligh. "Sparse code multiple access". In: *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)* (2013).

[NG19] Y. S. Nasir and D. Guo. "Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks". In: *EEE Journal on Selected Areas in Communications* 37.10 (2019), pp. 2239–2250.

[NNN19] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. "Deep Reinforcement Learning for Multi-Agent Systems: A Review of Challenges, Solutions and Applications". In: *IEEE Transactions on Cybernetics* (2019).

[Rap98] P. Rapajic. "Information capacity of the space division multiple access mobile communication system". In: *IEEE 5th Int. Symp. on Spread Spectrum Techniques and Applications* 3 (1998), pp. 946–950.

[Rei+21] R.-J. Reifert, A. A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx. "Ratesplitting multiple access in cache-aided cloud-radio access networks". In: *Frontiers in Commun. and Netw.* (2021).

[RHV13] V. Raghavan, S. Hanly, and V. Veeravalli. "Statistical beamforming on the grassmann manifold for the two-user broadcast channel". In: *IEEE Trans. on Info. Theory* 59(10) (2013), pp. 6464–6489.

[RN94] Gavin Adrian Rummery and Mahesan Niranjan. "On-line Q-learning using connectionist systems". In: 1994.

[Sai+13] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi. "Non-orthogonal multiple access (NOMA) for cellular future radio access". In: *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)* (2013).

[SB18] R. S. Sutton and A. G. Barto. "Reinforcement Learning: An Introduction". In: *Cambridge, MA, USA: MIT Press* (2018).

[SCO13] S.K. Sharma, S. Chatzinotas, and B. Ottersten. "SNR Estimation for Multidimensional Cognitive Receiver under Correlated Channel/Noise". In: *IEEE Trans. Wireless Commun.* 12(12) (2013), pp. 6392–6405.

[SH10] A. Sulyman and M. Hefnawi. "Performance evaluation of capacity-aware MIMO beamforming schemes in OFDM-SDMA systems". In: *IEEE Trans. on Commun.* 58(1) (2010), pp. 79–83.

[Sim+11] M. Simsek, A. Czylwik, A. Galindo-Serrano, and L. Giupponi. "Improved decentralized Q-learning algorithm for interference reduction in LTE-femtocells". In: *Proc. Wireless Adv.* (2011), pp. 138–143.

[Su+19] X. Su, L. Li, H. Yin, and P. Zhang. "Robust power- and rate-splittingbased transceiver design in k -user MISO SWIPT interference channel under imperfect CSIT". In: *IEEE Commun. Lett.* 23.3 (2019), pp. 514–517.

[Sun+15] Q. Sun, S. Han, C. L. I, and Z. Pan. "Energy efficiency optimization for fading MIMO non-orthogonal multiple access systems". In: *IEEE International Conference on Communications (ICC)* (2015).

[Sun+18] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos. "Learning to optimize: Training deep neural networks for interference management". In: *IEEE Trans. Signal Process.* 66.20 (2018), pp. 5438–5453.

[Tam+17] A. Tampuu, T. Matiisen, D. Kodelja, K. Kuzovkin I.and Korjus, J. Aru, and R. Vicente. "Multiagent cooperation and competition with deep reinforcement learning". In: *PloS One* 12.4 (2017).

[Tan95α] M. Tangemann. "Near-far effects in adaptive SDMA systems". In: *IEEE Int. Symp. PIMRC* 3 (1995), pp. 1293–1297.

[Tan95β] M. Tangemann. "Near-far effects in adaptive SDMA systems". In: *IEEE Int. Symp. PIMRC* 3 (1995), pp. 1293–1297.

[Ter+17] O. Tervo, A. Tolli, M. Juntti, and L. N. Tran. "Energy-efficient beam coordination strategies with rate-dependent processing power". In: *IEEE Transactions on Signal Processing* 65.22 (2017), pp. 6097–6112.

[Ter+18α] O. Tervo, L. Trant, S. Chatzinotas, B. Ottersten, and M. Juntti. "Multigroup multicast beamforming and antenna selection with ratesplitting in multicell systems". In: *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)* (2018).

[Ter+18β] O. Tervo, L. Trant, S. Chatzinotas, B. Ottersten, and M. Juntti. "Multigroup multicast beamforming and antenna selection with ratesplitting in multicell systems". In: *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)* (2018).

[TTJ15] O. Tervo, L. N. Tran, and M. Juntti. "Optimal energy-efficient transmit beamforming for multi-user MISO downlink". In: *IEEE Transactions on Signal Processing* 63.20 (2015), pp. 5574–5588.

[v80] 3GPP TR 36.913 v.8.0.0. *Requirements for further advancements for E-UTRA (LTE-Advanced).* URL: `http://www.3gpp.org`.

[Van+00] P. Vandenameele, L. Van der Perre, M. Engels, B. Gyselinckx, and H. De Man. "A combined OFDM/SDMA approach". In: *IEEE J. Sel. Areas in Commun.* 18(11) (2000), pp. 2312–2321.

[WD92] C.J.C.H. Watkins and P. Dayan. "Q-learning". In: *Machine Learning 8* 8(3-4) (1992), pp. 279–292. URL: `https://doi.org/10.1007/BF00992698`.

[Wie+10] Daan Wierstra, Alexander Forster, Jan Peters, and Jurgen Schmidhuber. "Recurrent Policy Gradients". In: *Logic Journal of the IGPL* 15.5 (2010), pp. 620–634.

[Wil92] Ronald J Williams. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". In: *Machine Learning* 8.3-4 (1992), pp. 229–256.

[Yan+18] Y. Yang, M. Pesavento, S. Chatzinotas, and B. Ottersten. "Energy efficiency optimization in MIMO interference channels: A successive pseudoconvex approximation approach". In: *arXiv preprint arXiv:1802.06750* (2018).

[YC21] L. Yin and B. Clerckx. "Rate-splitting multiple access for multigroup multicast and multibeam satellite systems". In: *IEEE Trans. Commun.* 69.2 (2021), pp. 976–990.

[YDC21] L. Yin, O. Dizdar, and B. Clerckx. "Rate-splitting multiple access for multigroup multicast cellular and satellite communications: PHY layer design and link-level simulations". In: *Proc. IEEE Int. Conf. Commun. (ICC) Workshop* (2021).

[YLJ18] H. Ye, G. Y. Li, and B.-H. Juang. "Power of deep learning for channel estimation and signal detection in OFDM systems". In: *IEEE Wireless Commun. Lett.* 7.1 (2018), pp. 114–117.

[YYC20] A. Z. Yalcin, M. Yuksel, and B. Clerckx. "Rate splitting for multi-group multicasting with a common message". In: *IEEE Trans. Veh. Technol.* 69.10 (2020), pp. 12 281–12 285.

[ZMC21] G. Zhou, Y. Mao, and B. Clerckx. "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff". In: *IEEE Trans. Wirel. Commun.* (2021), pp. 1–1.

[ZO95] P. Zetterberg and B. Ottersten. "The spectrum efficiency of a base station antenna array system for spatially selective transmission". In: *IEEE Trans. Vehicular Technol.* 44(3) (1995), pp. 651–660.

Γλωσσάριο

Απόδοση	Ξενόγλωσσος όρος
Πολλαπλή Πρόσβαση Διαίρεσης Ρυθμού	Rate Splitting Multiple Access (RSMA)
ανάθεση ισχύος	power allocation
Πληροφορίες Κατάστασης Καναλιού	Channel State Information (CSI)
Ενεργειακή Απόδοση	Energy Efficiency
Διαδίκτυο των Πραγμάτων	Internet of Things (IoT)
Σταθμισμένο Άθροισμα ρυθμών μετάδοσης	Weighted Sum Rate (WSR)
Συστ. Πολλαπλής Εισόδου-Μονής Εξόδου	Multiple Input-Single Output systems (MISO)
Σηματοθορυβικός Λόγος	Signal Noise Ratio (SNR)
Διαδοχική Κυρτή Προσέγγιση	Successive Convex Approximation (SCSA)
Συστ. Πολλαπλής Εισόδου-Πολλαπλής Εξόδου	Multiple Input- Multiple Output systems (MIMO)
Πολλαπλή Πρόσβαση Διαίρεσης Χώρου	Space Division Multiple Access (SDMA)
Μη Ορθογωνική Πολλαπλή Πρόσβαση	Non Orthogonal Multiple Access (NOMA)
Πολλαπλή Πρόσβαση Διαίρεσης Συχνότητας	Frequency Division Multiple Access (FDMA)
Πολλαπλή Πρόσβαση Ορθογωνικής Διαίρεσης Συχνότητας	Orthogonal Frequency Division Multiple Access
Ενισχυτική Μάθηση	Reinforcement learning (RL)
Βαθιά Ενισχυτική Μάθηση	Deep Reinforcement Learning (DRL)
Πολλαπλή Πρόσβαση Διαίρεσης Χρόνου	Time Division Multiple Access (TDMA)
Πολλαπλή Πρόσβαση Διαίρεσης Κώδικα	Code Division Multiple Access (CDMA)
Σταθμός Βάσης	Base Station (BS)
κατώφλι	threshold
ανερχόμενη ζεύξη	uplink
κατερχόμενη ζεύξη	downlink
από κοινού εκτίμηση καναλιού	joint channel estimation
Διαδοχική Ακύρωση Παρεμβολών	Successive Interference Cancellation (SIC)
βιβλίο κωδικών	codebook
προκωδικοποιητής	precoder

διάνυσμα διαμόρφωσης δέσμης	beamforming vector
Λευκός Προσθετικός Θόρυβος Γκάους	Additive White Gaussian Noise (AWGN)
Φυσικό στρώμα	Physical Layer (PHY)
τερματική κατάσταση	terminal state
Μαρκοβιανή Διαδικασία Απόφασης	Markov Decision Process (MDP)
Συνάρτηση Τιμής	Value Function
Συναρτησιακή Προσέγγιση Τιμής	Value Function Approximation
Αναζήτηση Πολιτικής	Policy Search
ανοδική/καθοδική κλίση	gradient ascent/descent
Συστήματα Πολλαπλών Πρακτόρων	Multi-Agent Systems (MASs)
Μερικώς Παρατηρήσιμη Διαδικασία Απόφασης Markov	Partially Observable Markov Decision Process
υπερπροσαρμογή	overfitting
κεντρική εκπαίδευση	centralized training
αποκεντρωμένη εκτέλεση	decentralized execution
διαμοιρασμός παραμέτρων	parameter sharing
Σταθμισμένο Μέσο Τετραγωνικό Σφάλμα	Weighted Mean Square Error (WMSE)
Δυναμικός Προγραμματισμός	Dynamic Programming
Συνελικτικό Νευρωνικό Δίκτυο	CNN