ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Ενισχυτική Μάθηση για τον Σχεδιασμό Τροχιών

Ονομ/νυμο: Σταμάτιος Ορφανός
Αριθμός Μητρώου: Ε17113
Επιβλέπων Καθηγητής: Βούρος Γεώργιος

Πειραιάς 2021

Περιεχόμενα

Στόχος Εργασίας - Μονοπρακτορικό Περιβάλλον

Ο βασικός στόχος της εργασίας για το μονοπρακτορικό περιβάλλον είναι η εύρεση του βέλτιστου μονοπατιού προς τον κόμβο στόχο αποφεύγοντας τα εμπόδια του περιβάλλοντος σε ένα πεπερασμένο σύνολο επεισοδίων.

Βασικό κριτήριο για την αξιολόγηση της απόδοσης είναι η ταχύτητα με την οποία ο πράκτορας έβρισκε το βέλτιστο μονοπάτι. Έτσι δημιουργήσαμε τις εξής τρεις περιπτώσεις για την διεξαγωγή πειραμάτων:

1. Περιβάλλον με σταθερά εμπόδια
2. Περιβάλλον με εμπόδια που μετακινούνται κάθε 1.000 επεισόδια
3. Περιβάλλον με εμπόδια που μετακινούνται κάθε 10.000 επεισόδια

Στόχος Εργασίας - Πολυπρακτορικό Περιβάλλον

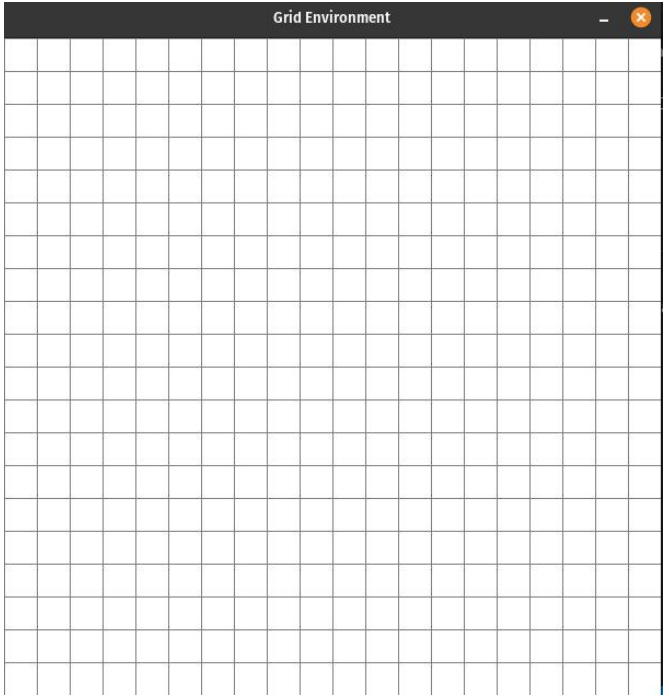
Ο βασικός στόχος της εργασίας για το πολυπρακτορικό περιβάλλον είναι η εύρεση των βέλτιστων μονοπατιών προς τους κόμβους στόχους για κάθε πράκτορα, ενώ παράλληλα ο κάθε πράκτορας θα πρέπει να αποφεύγει τον άλλο. Ομοίως βασικά κριτήρια αποτελούν:

1. Η ταχύτητα με την οποία ο κάθε πράκτορας μαθαίνει το περιβάλλον
2. Η ακρίβεια της πολιτικής που αναπτύχθηκε για το περιβάλλον

Διατύπωση Προβλήματος - Περιβάλλον

Το περιβάλλον της εφαρμογής είναι ένα πλέγμα με τα παρακάτω χαρακτηριστικά:

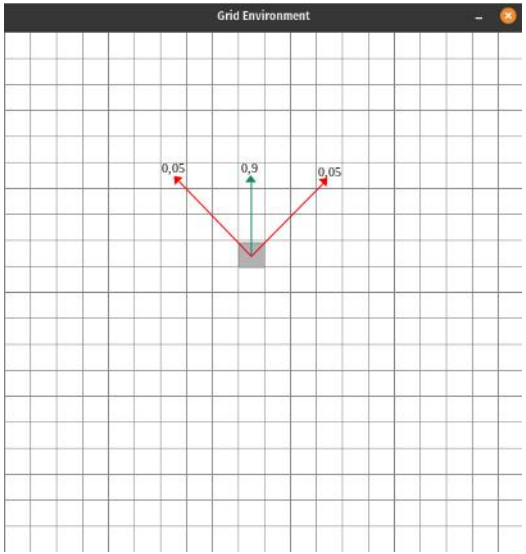
1. Ανοικτό περιβάλλον
2. Μέγεθος 20x20
3. Περιλαμβάνει στοιχεία στοχαστικότητας
4. Η κίνηση μέσα σε αυτό δίνει μικρή αρνητική ανταμοιβή

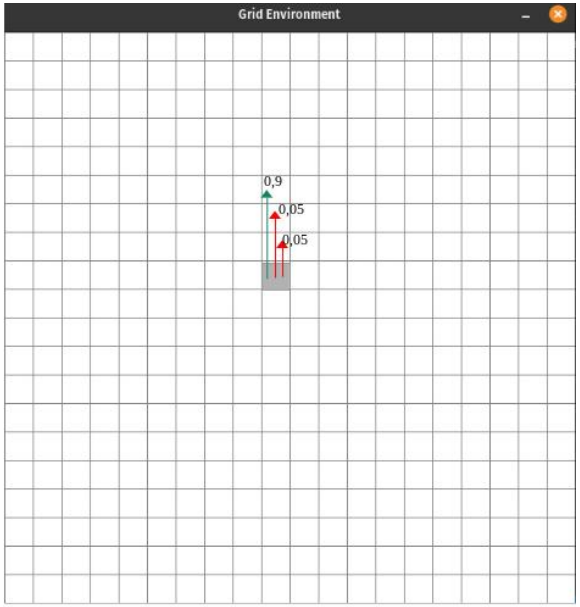


Διατύπωση Προβλήματος - Στοχαστικότητα

Η στοχαστικότητα του περιβάλλοντος αποτελείται από δύο παραμέτρους:

1. Στοχαστικότητα Κατεύθυνσης
2. Στοχαστικότητα Ταχύτητας

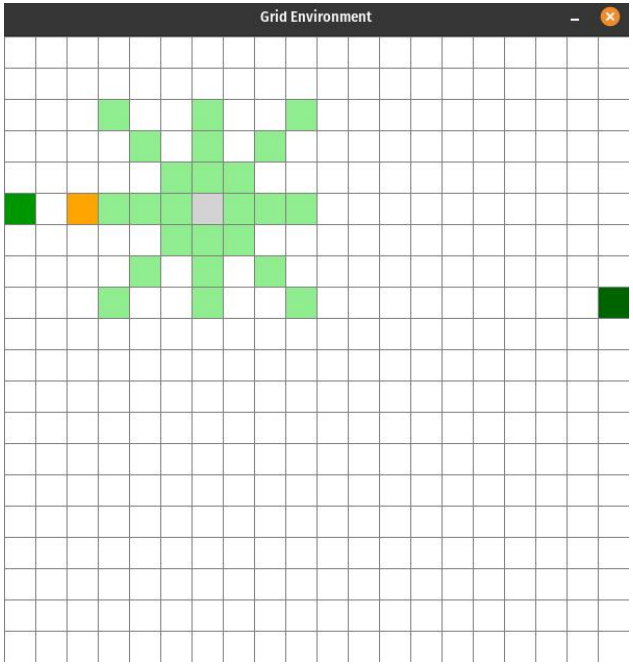




Διατύπωση Προβλήματος - Πράκτορας

Ένας πράκτορας θεωρείται μια οντότητα του περιβάλλοντος που διαθέτει τα παρακάτω χαρακτηριστικά:

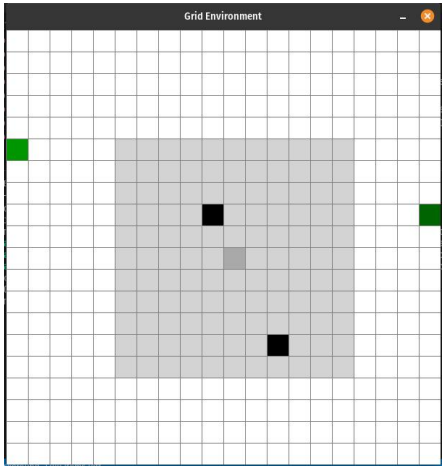
1. Χαρακτηριστικά θέσης στο περιβάλλον
2. Μια λίστα ανταμοιβών για κάθε επεισόδιο
3. Μια λίστα ανταμοιβών για όλα τα επεισόδια
4. Ένα πίνακα Q-table
5. Ένα πίνακα Elegibility-table
6. Ένα heatmap
7. Μια λίστα με τις τιμές αέρα για κάθε βήμα
8. Ένα μονοπάτι για κάθε επεισόδιο
9. Ένα κόμβο εκκίνησης και ένα κόμβο τερματισμού

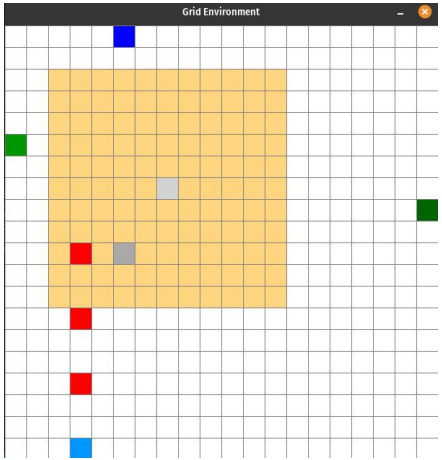


Διατύπωση Προβλήματος - Πράκτορας

Ένας πράκτορας διαθέτει τις εξής δυνατότητες:

1. Τυχαία κίνηση στο περιβάλλον
2. Κίνηση με βάση την πολιτική στο περιβάλλον
3. Ενημέρωση και χρήση των πινάκων Q-table και Elegibility-table
4. Δημιουργία του heatmap κίνησης
5. Έλεγχος κίνησης στο περιβάλλον σε περίπτωση εμποδίων
6. Έλεγχος και λήψη των ανταμοιβών του περιβάλλοντος





Προτεινόμενη Λύση - E-Greedy

Η E-Greedy Strategy είναι μια μέθοδος εξερεύνησης-εκμετάλλευσης του περιβάλλοντος όπου:

- Η άπληστη ενέργεια αντιπροσωπεύει αυτό που θα πρότεινε η πολιτική μας για βέλτιστα αποτελέσματα και θα ακολουθήσουμε αυτήν την πολιτική με πιθανότητα 1 − ε.
- Ωστόσο, θα καταφύγουμε σε τυχαία επιλογή με πιθανότητα ε, επιτρέποντάς μας μερικές φορές να κάνουμε μη βέλτιστες επιλογές για χάρη της εξερεύνησης. Συχνά, αυτές οι τυχαίες κινήσεις εξερευνήσεις θα ανακαλύψουν μια πιο πολύτιμη κατάσταση από αυτήν που γνώριζε η πολιτική μας και έτσι μπορούμε να ενημερώσουμε ανάλογα την πολιτική μας.

Ένα από τα βασικά χαρακτηριστικά της E-Greedy μεθόδου είναι ο συνδυασμός τυχαίων κινήσεων και άπλειστων κινήσεων παρέχοντας μια σταδιακή και ομαλή μετάβαση μεταξύ των δυο συμπεριφορών.

Προτεινόμενη Λύση - Monte-Carlo-Temporal Difference Hybrid

Η Monte-Carlo-Temporal Difference Hybrid είναι μια μέθοδος εξερεύνησης-εκμετάλλευσης του περιβάλλοντος όπου:

- Στο στάδιο της εξερεύνησης ο πράκτορας εκτελεί τυχαίες κινήσεις στο περιβάλλον προκειμένου να μάθει το περιβάλλον με γρήγορο ρυθμό.
- Στο στάδιο της εκμετάλλευσης ο πράκτορας εκτελεί μόνο κινήσεις βάση της πολιτικής που έχει αναπτύξει στο στάδιο της εξερεύνησης.

Σε αυτή την περίπτωση το στάδιο εξερεύνησης και το στάδιο εκμετάλλευσης είναι διαφορετικά και δεν υπάρχει κάποιος συνδυασμός εξερεύνησης-εκμετάλλευσης.

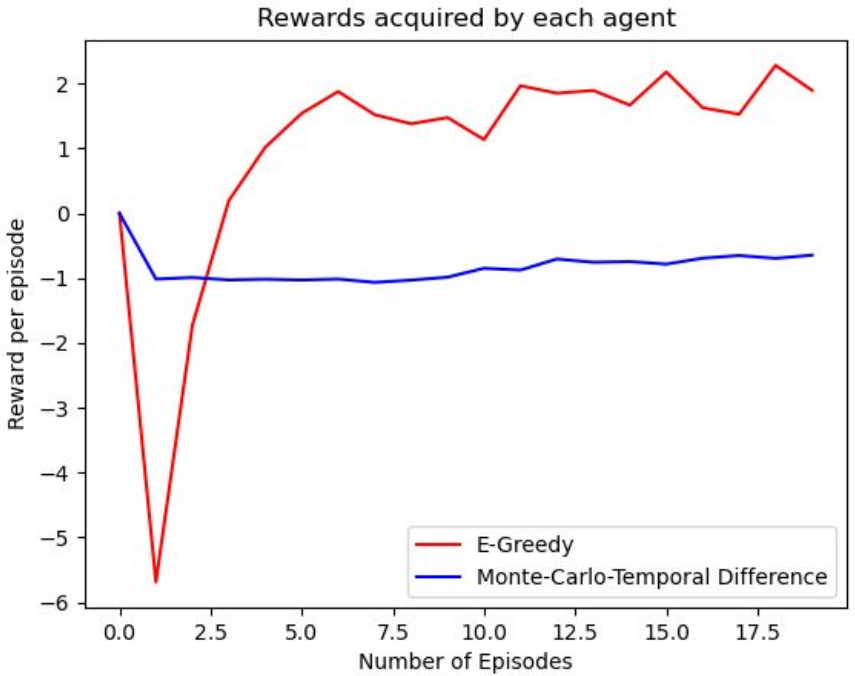
Πειραματικές Μελέτες

Για την διεξαγωγή των πειραματικών μελετών είχαμε τις εξής παραμέτρους:

1. **Αριθμός επεισοδίων 100.000**
2. **Ρυθμός Μάθησης α = 0.001**
3. **Παράγοντας Προεξόφλησης γ = 1**
4. **Παράμετρος λ = 0.65**

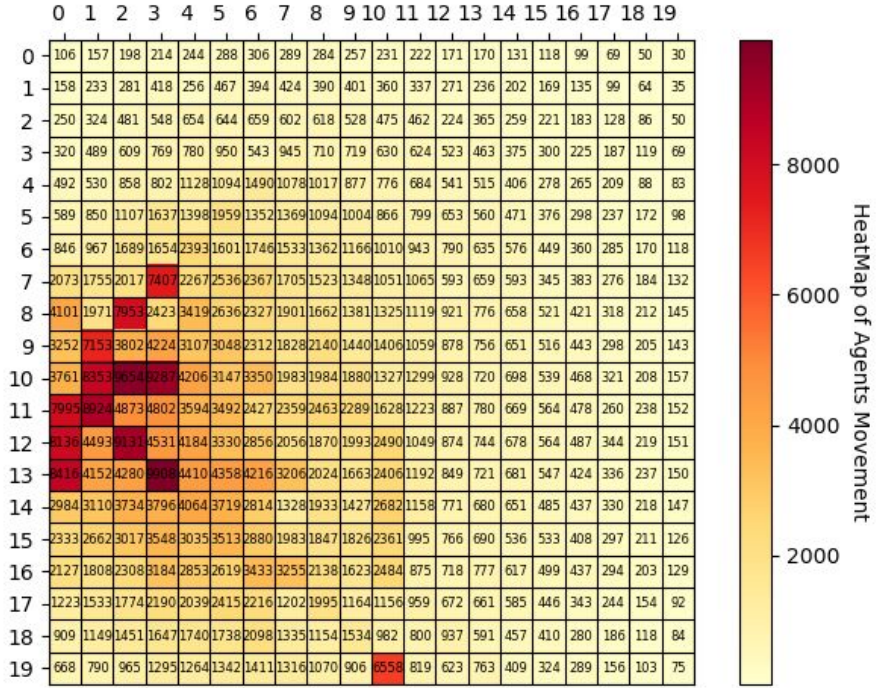
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Σταθερά Εμπόδια



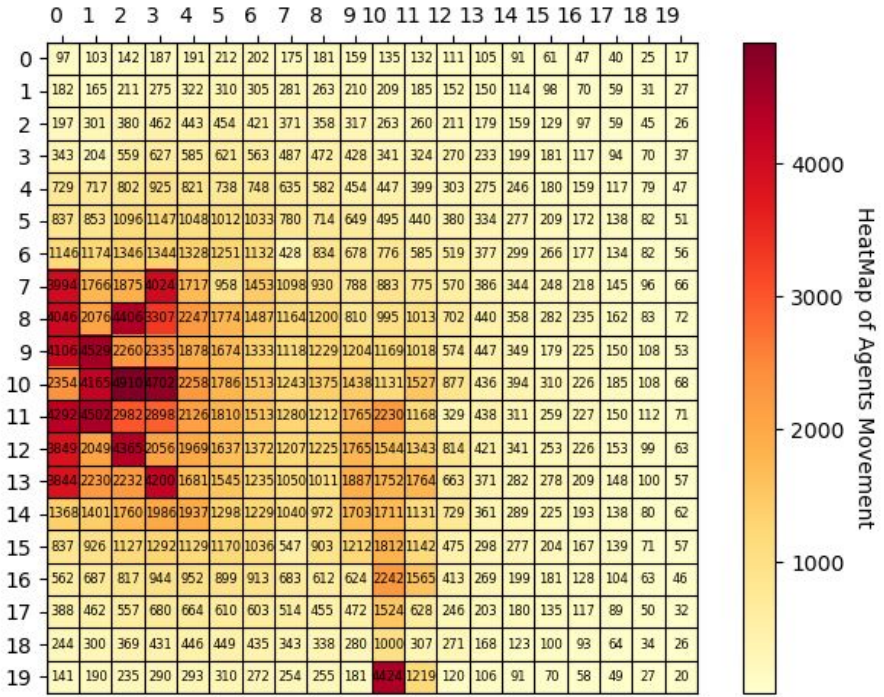
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Σταθερά Εμπόδια



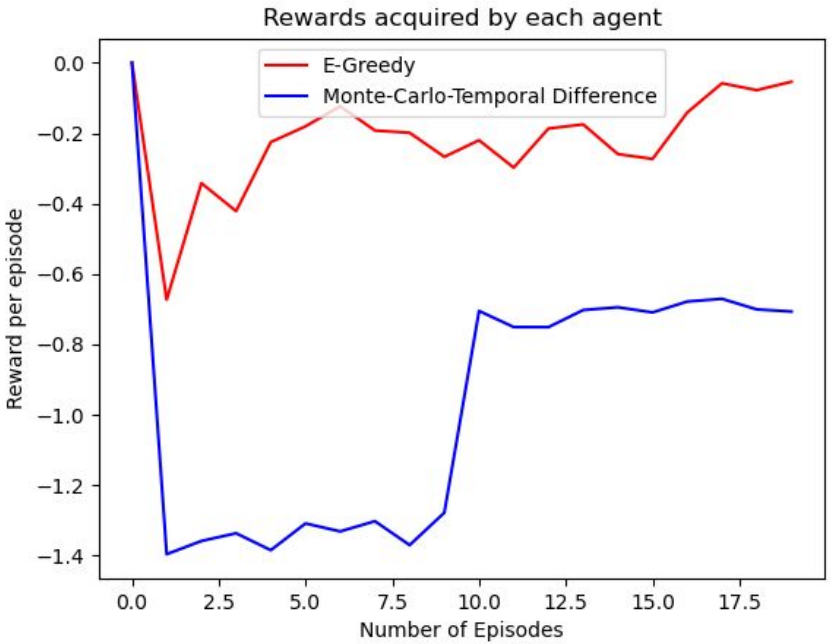
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Σταθερά Εμπόδια



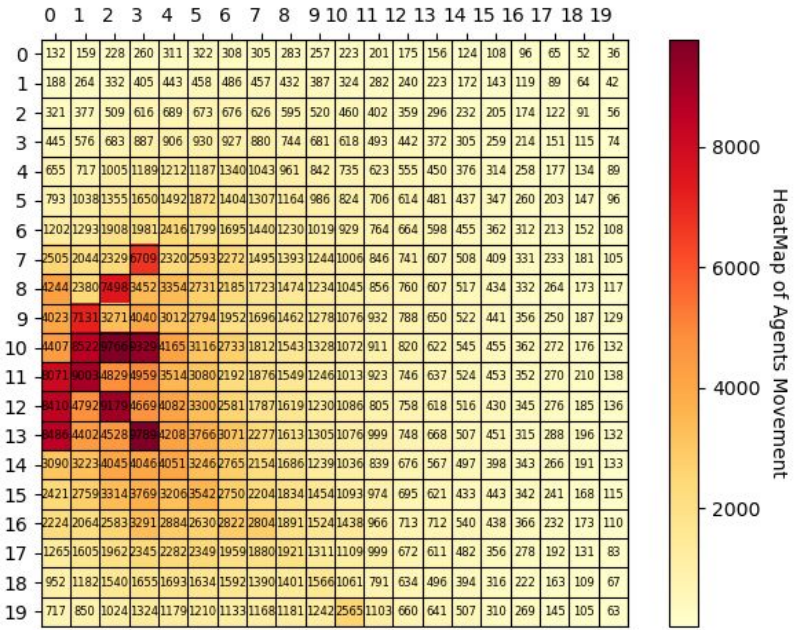
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Εμπόδια που κινούνται ανά 1.000 επεισόδια



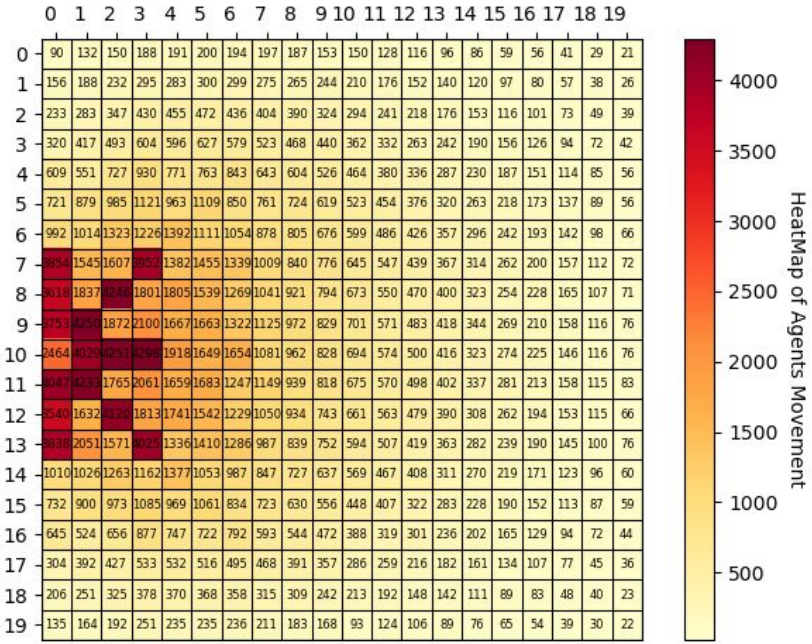
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Εμπόδια που κινούνται ανά 1.000 επεισόδια



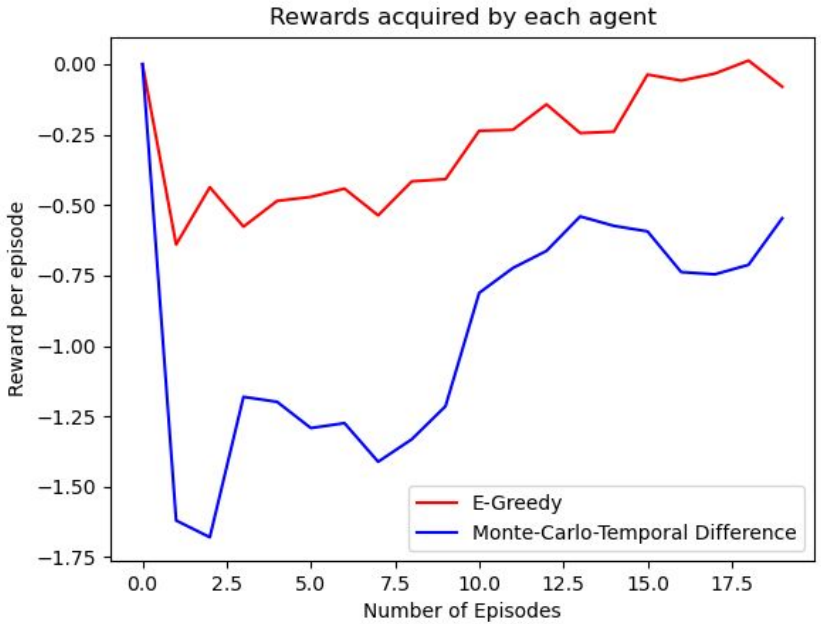
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Εμπόδια που κινούνται ανά 1.000 επεισόδια



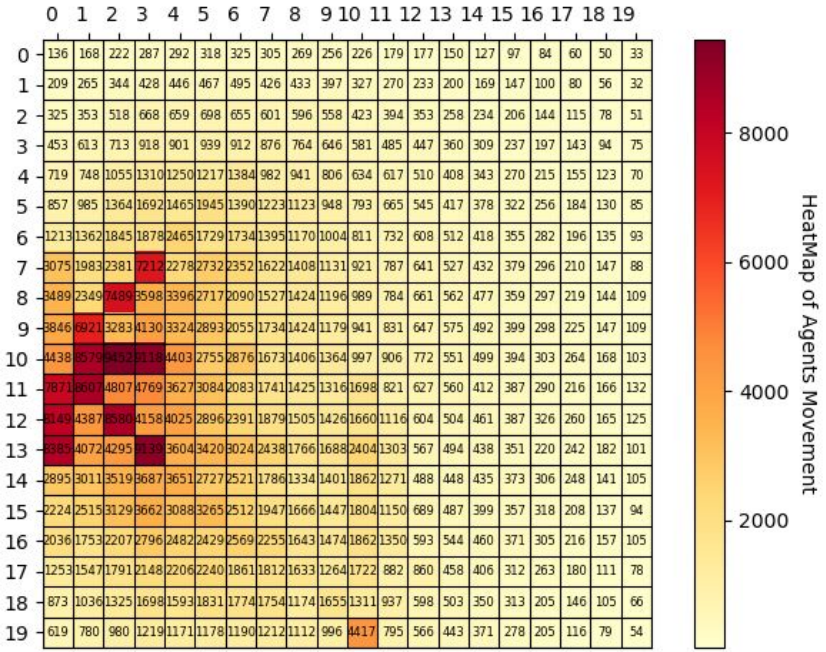
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Εμπόδια που κινούνται ανά 10.000 επεισόδια



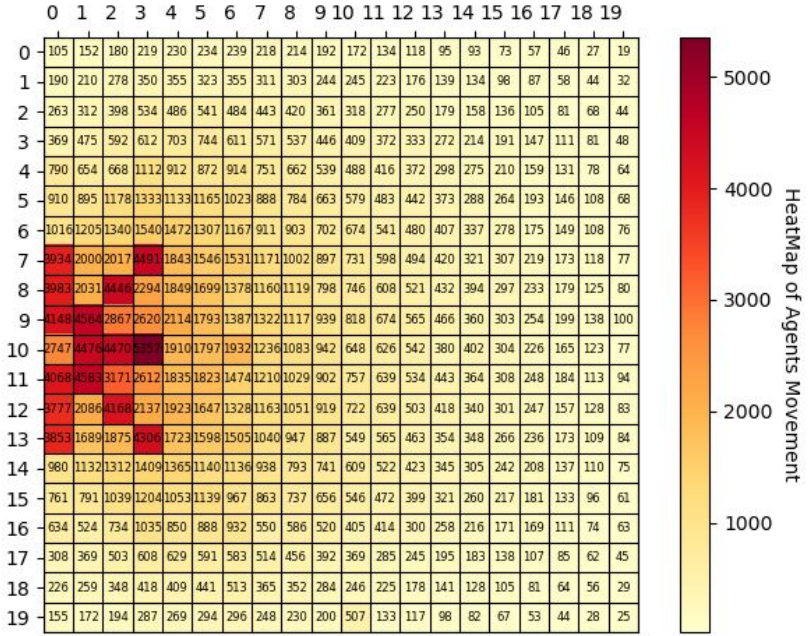
Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

Εμπόδια που κινούνται ανά 10.000 επεισόδια

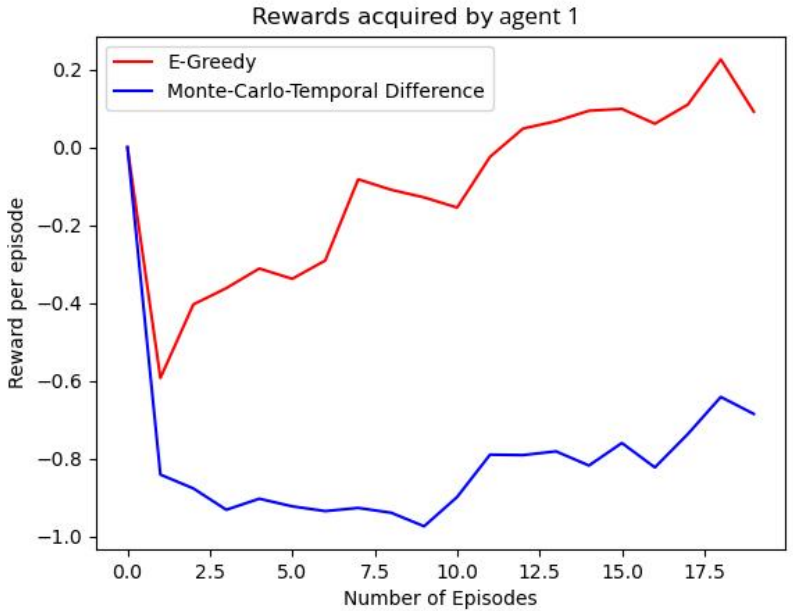


Πειραματικές Μελέτες - Μονοπρακτορικό Περιβάλλον

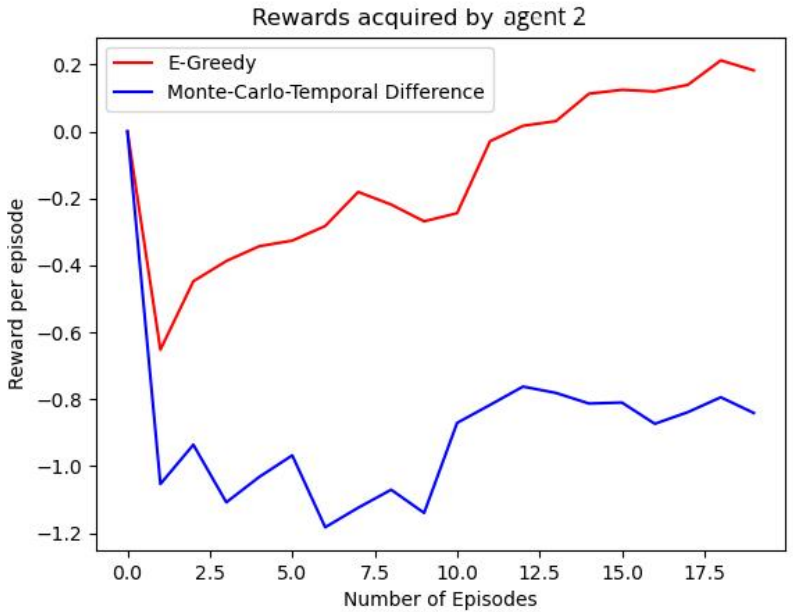
Εμπόδια που κινούνται ανά 10.000 επεισόδια



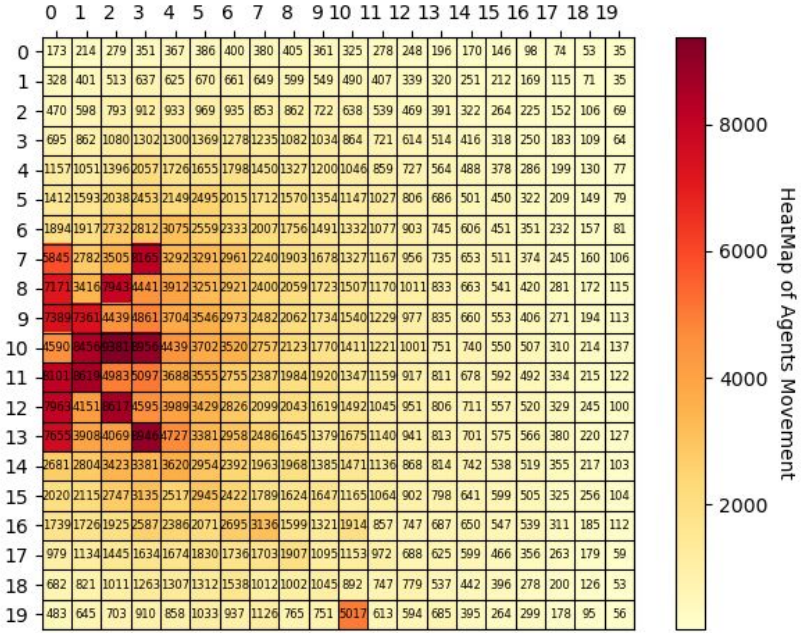
Πειραματικές Μελέτες - Πολυπρακτορικό Περιβάλλον



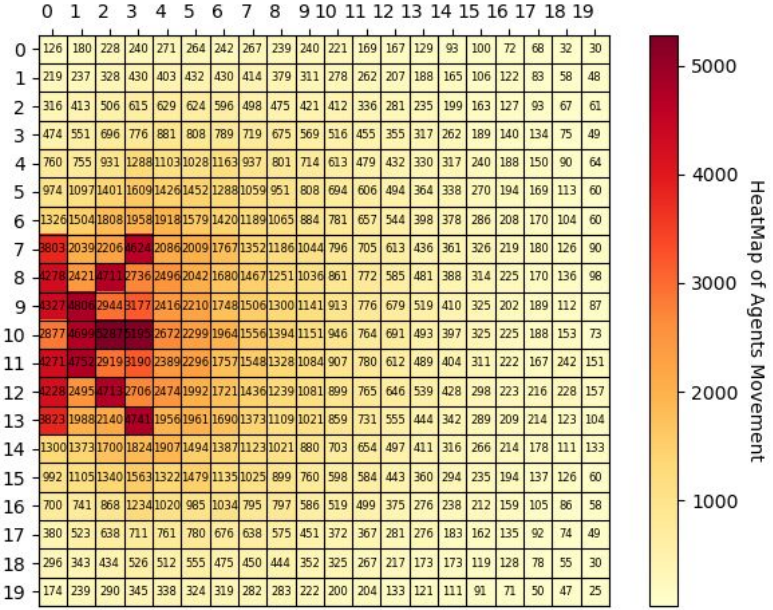
Πειραματικές Μελέτες - Πολυπρακτορικό Περιβάλλον



Πειραματικές Μελέτες - Πολυπρακτορικό Περιβάλλον

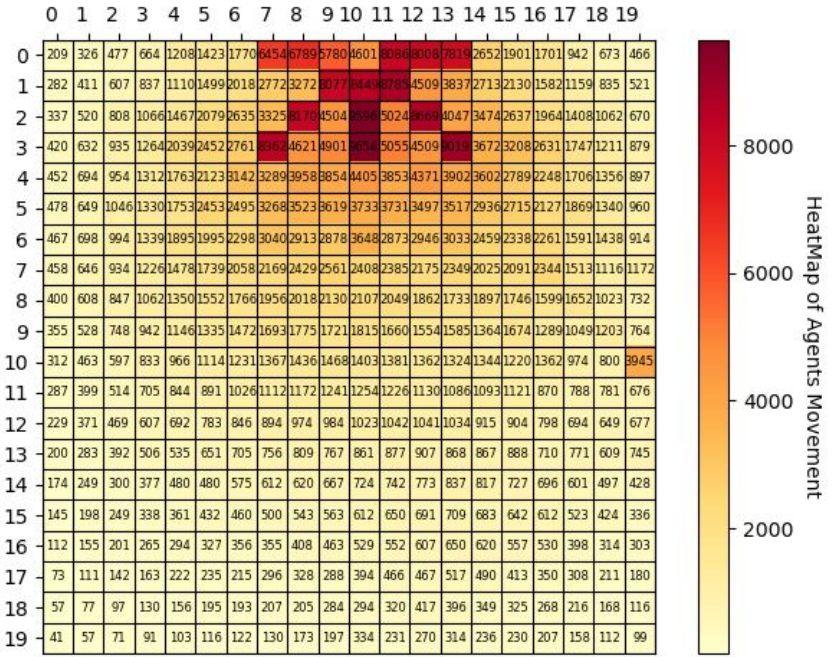


E-Greedy Strategy

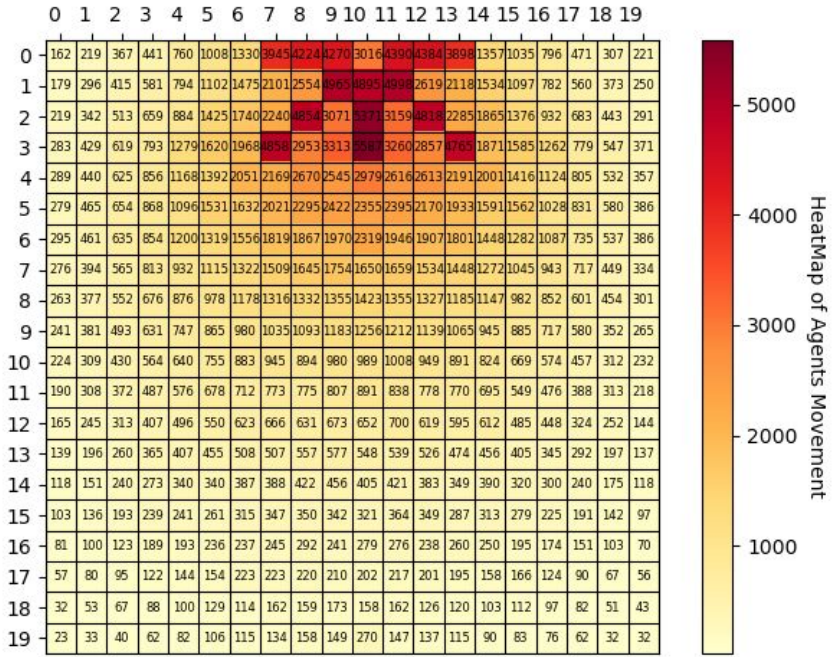


Monte-Carlo-Temporal Difference Hybrid

Πειραματικές Μελέτες - Πολυπρακτορικό Περιβάλλον



E-Greedy Strategy



Monte-Carlo-Temporal Difference Hybrid

Συμπεράσματα - Μονοπρακτορικό Περιβάλλον

- Η μέθοδος E-Greedy Strategy κατάφερε στις 2 από τις 3 περιπτώσεις να προσεγγίσει τον κόμβο στόχο, αποδεικνύοντας την αποτελεσματικότητα της σε δυναμικά περιβάλλοντα υπογραμμίζοντας την σημασία της σταδιακής εναλλαγής από εξερεύνηση σε εκμετάλλευση.

- Η μέθοδος Monte-Carlo-Temporal Difference Hybrid είχε επιτυχία όταν το περιβάλλον δεν άλλαζε κατά τη διάρκεια του πειράματος με μεγάλες θετικές ανταμοιβές σε ορισμένα πειράματα, ενώ είχε αποτυχία όταν το περιβάλλον εκμετάλλευσης ήταν αρκετά διαφορετικό σε σχέση με το αντίστοιχο εξερεύνησης.

Συμπεράσματα - Πολυπρακτορικό Περιβάλλον

- Η μέθοδος E-Greedy Strategy ήταν αποτελεσματική για τους δυο πράκτορες, οι οποίοι κατάφεραν να βρουν ένα σχεδόν βέλτιστο μονοπάτι, αφού έλαβαν κατά μέσο όρο μικρές θετικές ανταμοιβές.

- Η μέθοδος Monte-Carlo-Temporal Difference Hybrid δεν κατάφερε να λύσει το πρόβλημα. Σε αυτή την περίπτωση η γρηγορότερη μετάδοση της γνώσης για το περιβάλλον αποτέλεσε πρόβλημα, διότι από τα πρώτα επεισόδια η μεγάλη αρνητική ανταμοιβή είχε αποτυπωθεί στον πίνακα Q-table, λόγω της συνύπαρξης των δυο πρακτόρων σε κοντινή περιοχή.

Demo

Σας ευχαριστώ για την προσοχή σας