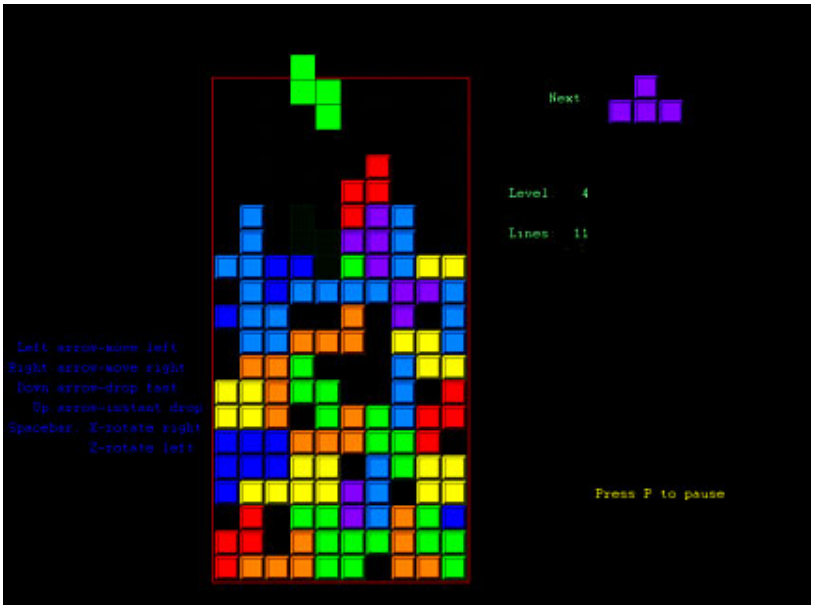
TETRIS με Ενισχυτική Μάθηση



Στέφανος Βαρότσης

TETRIS με Ενισχυτική Μάθηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Στέφανος Βαρότσης

A.M: 2004010004, Email: svarotsis@isc.tuc.gr

Εξεταστική Επιτροπή:

Ν. Βλάσσης, Επίκουρος Καθηγητής, Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Πολυτεχνείο Κρήτης (Επιβλέπων)

Μ. Λαγουδάκης, Επίκουρος Καθηγητής, Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης

Ι. Νικολός, Επίκουρος Καθηγητής, Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Πολυτεχνείο Κρήτης

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ
ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Εικόνα εξωφύλλου: το παιγνίδι Tetris

TETRIS με Ενισχυτική Μάθηση

Περίληψη

Στην παρούσα διπλωματική παρουσιάζουμε την εφαρμογή της Ενισχυτικής Μάθησης στο ηλεκτρονικό παιχνίδι Tetris ώστε να επιτευχθεί μια γρήγορη και ”έξυπνη” στρατηγική παιξίματος, δηλαδή ο παίκτης να παίζει γρήγορα και αποτελεσματικά έναντι σε δύσκολες συνθήκες παιξίματος.Επίσης, εδώ επεξηγείται το τι ακριβώς είναι η Ενισχυτική Μάθηση, πού χρησιμεύει και γιατί και με ποιο τρόπο το κάνει αυτό. Έγινε μια σύγκριση ενός αλγορίθμου Ενισχυτικής Μάθησης που ονομάζεται cross entropy και μιας τυχαίας στρατηγικής παιξίματος και αναλύθηκαν τα αποτελέσματα αυτής της σύγκρισης.

Πρόλογος

Ευχαριστίες

Πρώτα από όλα, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κύριο Νίκο Βλάσση για τις πολύτιμες γνώσεις που μου πρόσφερε. Επίσης, είναι ο άνθρωπος που με μύησε στον κόσμο της ρομποτικής και της Ενισχυτικής Μάθησης, που με έμαθε να σκέφτομαι ορθά και να εργάζομαι συστηματικά και με ζήλο που πάντα με καθοδηγούσε και βοηθούσε σε οποιοδήποτε πρόβλημα αντιμετώπιζα. Οι συζητήσεις μας ήταν ένα σωτήριο εφόδιο για την επιτυχή ολοκλήρωση της διπλωματικής μου εργασίας.

Έπειτα, θα ήθελα να δώσω ένα ξεχωριστό ευχαριστώ στον Γιώργο Κόντε για τις πάρα πολλές ώρες που μου διέθεσε και τις γνώσεις και συμβουλές του πάνω στην Ενισχυτική Μάθηση και τον κώδικα της διπλωματικής.Επίσης, και για το γεγονός ότι ήταν πάντα διαθέσιμος όταν τον χρειαζόμουν και η παρέα του ήταν και είναι ακόμα πάρα πολύ ευχάριστη.

Ένα άτομο το οποίο επιβάλλεται να ευχαριστήσω είναι ο κολλητός μου ο Walid Soulakis ο οποίος ήταν πάντα διατεθιμένος να με βοηθήσει και με τον οποίο πέρασα ατελείωτες ώρες στο εργαστήριο (και θα περάσουμε από ότι φαίνεται άλλες τόσες!!!).Επιπρόσθετα, ευχαριστώ και τους υπόλοιπους κολλητούς (Νίκο, Ευτύχη, Γιάννη, Δημήτρη Βακόνδιο) για τις υπέροχες στιγμές που περάσαμε και θα περάσουμε μαζί. Τέλος, θα ήθελα να ευχαριστήσω πολύ την οικογένειά μου για την ηθική και υλική τους υποστήριξη όλα αυτά τα χρόνια.

Περιεχόμενα

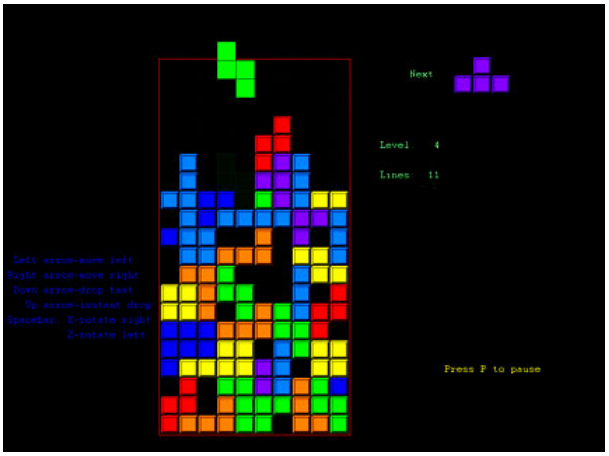
Κεφάλαιο 1

Εισαγωγή

1.1 Το παιχνίδι Tetris και οι κανόνες που το διέπουν

Το Tetris είναι ένα ηλεκτρονικό παιχνίδι το οποίο σχεδιάστηκε και προγραμματίστηκε το 1985 από τον Alexey Pajitnov και από τότε έγινε ένα από τα πιο δημοφιλή ηλεκτρονικά παιχνίδια παγκοσμίως. Στο παιχνίδι αυτό μια ψευδοτυχαία ακολουθία τετρόνιμων (μερικές φορές αποκαλούμενα και "τετράδες" σε παλαιότερες εκδόσεις)-σχήματα που αποτελούνται από τέσσερα τετράγωνα τούβλα- πέφτουν ένα ένα διανύοντας τον πίνακα παιξίματος, το περιβάλλον του παιχνιδιού δηλαδή, από πάνω προς τα κάτω.

Ο σκοπός του παιχνιδιού είναι να χειριστεί ο παίκτης τα τετρόνιμα αυτά με την μετακίνησή τους αριστερά ή δεξιά ή/και με την περιστροφή τους αριστερόστροφα ή δεξιόστροφα(κατά πολλαπλάσια των 90 μοιρών), με στόχο να δημιουργηθεί μια οριζόντια γραμμή τούβλων δίχως κενά. Όταν δημιουργηθεί μια τέτοια γραμμή, τότε αυτή διαγράφεται και τα τούβλα που βρίσκονται από πάνω της εάν υπάρχουν πέφτουν τόσο όσο και το πλάτος της γραμμής αυτής. Άρα,με την διαγραφή κάθε σειράς τα υπερκείμενα τούβλα θα μετατοπιστούν προς τα κάτω καλύπτοντας την θέση της. Το παιχνίδι τερματίζεται όταν ο παίκτης βρίσκει "ταβάνι", δηλαδή όταν η στοίβα των τετρόνιμων φτάσει το πάνω



Σχήμα 1.1: Το TETRIS είναι ένα πολύ ωραίο παιγνίδι.

Σχήμα 1.2: Το TETRIS βρίσκει ”ταβάνι”.

μέρος του πίνακα παιξίματος και δεν μπορούν να εισαχθούν άλλα. Οι τύποι των κομματιών που χρησιμοποιούνται είναι ”Z”, ”T”, ”I”, ”S”, ”O”, ”J”, ”L”, δηλαδή τα σχήματα των τετρόνιμων μοιάζουν με τα γράμματα αυτά. Οι τύποι αυτοί παρουσιάζονται στην παρακάτω εικόνα :

Αν και έχει απλούς κανόνες, το να παίξει κανείς το παιχνίδι αυτό καλά απαιτεί μια πολύπλοκη στρατηγική παιξίματος και πολύ εμπειρία. Επιπλέον, οι Demaine et al. έχουν δείξει ότι το Tetris είναι δύσκολο και με την μαθηματική έννοια [Demaine et al., 2003]: η εύρεση της βέλτιστης στρατηγικής είναι ένα πρόβλημα NP-complete ακόμα και αν η ακολουθία των τετρόνιμων είναι γνωστή εκ των προτέρων. Αυτές οι ιδιότητες καθιστούν το Tetris το οποίο είναι ένα δύσκολο συνδυαστικό πρόβλημα και συγχρόνως ένα ενδιαφέρον πρόβλημα πειραματισμού για την βελτίωση της ταχύτητας επίλυσης των αλγορίθμων ώστε να βρεθεί ο ταχύτερος για πειραματικές εφαρμογές και για εφαρμογές αλγορίθμων Ενισχυτικής Μάθησης και γενικότερα μηχανικής μάθησης . Οι αλγόριθμοι Ενισχυτικής Μάθησης είναι αρκετά αποτελεσματικοί στην επίλυση μιας ποικιλίας πολύπλοκων προβλημάτων διαδοχικών αποφάσεων. Παρ’όλα αυτά, οι προσεγγίσεις της Ενισχυτικής Μάθησης στο Tetris για την βέλτιστη λύση έχουν παραδόξως δείξει μη ικανοποιητική απόδοση. Με την εργασία αυτή θα επιχειρηθεί η εφαρμογή της μεθόδου Ενισχυτικής Μάθησης cross entropy για την βελτίωση της απόδοσης του παίκτη. Στην εργασία αυτή εξετάζεται η εφαρμογή αλγόριθμου Ενισχυτικής Μάθησης στο παιχνίδι Tetris για την εύρεση της βέλτιστης στρατηγικής παιξίματος με την σύγκλιση στις βέλτιστες τιμές των παραμέτρων που παιχνιδιού. Επίσης, θα συγκριθεί η αποτελεσματικότητά της ως προς μια τυχαία πολιτική. Θα ακολουθηθούν οι κανόνες που δόθηκαν από τον διαγωνισμό RL competition 2009 οι οποίοι πέρα από τους κανόνες του κλασικού παιχνιδιού που προαναφέρθηκαν στηρίζονται με κάποιες διαφορές και στην ειδική έκδοση του Van Roy (1995) για το Tetris και είναι οι εξής :

- ο παίκτης έχει τη δυνατότητα να περιστρέψει ή να μετακινήσει το τούβλο κατά ένα διάστημα

μεγέθους ενός τούβλου καθώς αυτό πέφτει (όπως συμβαίνει και στην εμπορική έκδοση του ηλεκτρονικού παιχνιδιού). Στην ειδική έκδοση ο παίκτης επιλέγει μια τελική θέση και προσανατολισμό για κάθε τούβλο όταν αυτό πρωτοεμφανίζεται.

- Με την ταυτόχρονη διαγραφή τεσσάρων γραμμών αποκτάται περισσότερη ανταμοιβή από το εάν διαγράφονταν οι γραμμές αυτές μία προς μία.
- Η κατανομή των τύπων των τετρόνιμων που εμφανίζονται δεν είναι ομοιόμορφη. Για παράδειγμα, σε μερικά παιξίματα μπορεί να υπάρχει πλεόνασμα κομματιών ”Z” ενώ σε άλλα να υπάρχει έλλειψη. Επίσης η πιθανότητα να εμφανιστεί ένα επιθυμητό τετρόνιμο κατά τη διάρκεια του παιχνιδιού είναι αντιστρόφως ανάλογη του βαθμού επιθυμίας από τον παίκτη για το τετρόνιμο αυτό.

Κεφάλαιο 2

Ενισχυτική μάθηση

Στο κεφάλαιο αυτό θα δοθεί μια εισαγωγή στον ορισμό της Ενισχυτικής Μάθησης, στα συστατικά από τα οποία αποτελείται, στο πώς τα συστατικά αυτά συνδέονται μεταξύ τους καθώς και μια περιγραφή του θεωρητικού μέρους του υπό μελέτη προβλήματος.

2.1 Ορισμός Ενισχυτικής Μάθησης

Η βασική ιδέα της **Ενισχυτικής Μάθησης** είναι η περίπτωση ενός συστήματος που προσπαθώντας να επιτύχει ένα στόχο προσαρμόζει την συμπεριφορά του στις αλλαγές του περιβάλλοντός του ώστε να μεγιστοποιήσει ένα ειδικό σήμα που λαμβάνει από αυτό. Το σήμα αυτό ο παίκτης το αντιλαμβάνεται ως μια ανταμοιβή για την ορθότητα της ενέργειας που επέλεξε. Με άλλα λόγια, με την **Ενισχυτική Μάθηση**, όπως και σε άλλη μάθηση, η εκμάθηση επιτυγχάνεται μέσω διεπαφής του παίκτη και του περιβάλλοντός του. Η πορεία του παίκτη προς τον στόχο γίνεται με το πέρασμά του από διάφορες καταστάσεις του συστήματος μέσω επιλεγμένων ενεργειών του. Ανάλογα με τις ενέργειες που θα επιλέξει θα μεταβεί και σε άλλη κατάσταση του συστήματος.

Τα δύο σημαντικότερα χαρακτηριστικά της **Ενισχυτικής Μάθησης** είναι η αναζήτηση δοκιμής και σφάλματος και οι ανταμοιβές σε ύστερο χρόνο, ανταμοιβές δηλαδή που θα πάρει ο παίκτης εάν εκτελέσει μια συγκεκριμένη ενέργεια. Ένα παράδειγμα για την καλύτερη κατανόηση των όσων αναφέρθηκαν παραπάνω είναι το εξής: Στο υπό μελέτη ηλεκτρονικό παιχνίδι Tetris ο παίκτης πρέπει να καταλάβει, μέσω της εμπειρίας του παιξίματός του έναντι του υπολογιστή,τον αντίπαλό του, ποιες κινήσεις θα του αποφέρουν μεγαλύτερη ανταμοιβή, καλύπτοντας δηλαδή τα κενά στις γραμμές των τούβλων που σχηματίζονται με τον καλύτερο δυνατό τρόπο. Στην περίπτωσή μας η αναζήτηση δοκιμής και σφάλματος θα γινόταν με αρχικά τυχαίες τοποθετήσεις κομματιών και ανάλογα με το αποτέλεσμα θα γινόταν και η αντίστοιχη εκμάθηση. Οι ανταμοιβές σε ύστερο χρόνο έχουν την λογική ότι ο παίκτης δίδει περισσότερη σημασία μόνο στην μεγιστοποίηση των ανταμοιβών σε βάθος χρόνου, με άλλα λόγια δεν τον ενδιαφέρουν τόσο οι πρόσκαιρες ανταμοιβές.

Επίσης, σημασία πρέπει να δοθεί και στο γεγονός ότι η Ενισχυτική Μάθηση καθορίζεται από τον χαρακτηρισμό προβλημάτων μάθησης και όχι χαρακτηρίζοντας μεθόδους μάθησης. Επιπρόσθετα, η Ενισχυτική Μάθηση δεν πρέπει να συγχέεται με την επιβλεπόμενη μάθηση, κατά την οποία ένας επιβλέπων δίδει στον παίκτη παραδείγματα ώστε να μάθει και να αποκτήσει εμπειρία. Για παράδειγμα, στο Tetris ο επιβλέπων θα έδειχνε στον παίκτη ποιες είναι οι σωστές τοποθετήσεις των κομματιών σε κάθε χρονικά μικρό παίξιμο του παιχνιδιού. Μόνο στις περιπτώσεις όπου κρίνεται ότι συγκεκριμένες ενέργειες του παίκτη είναι ιδιαίτερα κρίσιμες εφαρμόζεται η επιβλεπόμενη μάθηση.

Μια άλλη διαφορά των δύο αυτών μεθόδων μάθησης συνίσταται στο ότι η Ενισχυτική Μάθηση περιέχει αξιολογική ανάδραση που υποδεικνύει το πόσο καλή είναι μια ενέργεια αλλά όχι το εάν είναι η καλύτερη ή χειρότερη δυνατή. Από την άλλη, η επιβλεπόμενη μάθηση χρησιμοποιεί ανάδραση καθοδήγησης που υποδεικνύει την σωστή ενέργεια που πρέπει να γίνει ανεξάρτητα από την ενέργεια η οποία πραγματικά έγινε.

Ένα βασικό ζήτημα της Ενισχυτικής Μάθησης που αποτελεί και μια μεγάλη πρόκληση είναι η σωστή ισορροπία που πρέπει να διατηρηθεί μεταξύ εξερεύνησης και εκμετάλλευσης του χώρου αναζήτησης της λύσης. Ο παίκτης μέσω της εκμετάλλευσης του χώρου αναζήτησης αναζητά την ενέργεια αυτή που θα του προσδώσει την περισσότερη ανταμοιβή ανάμεσα σε αυτές που έχει ήδη εξερευνήσει. Όμως, με αυτόν τον τρόπο ρισκάρει να μην βρει μια πιθανή ενέργεια που θα του δώσει ακόμα περισσότερη ανταμοιβή. Επομένως, πρέπει **και** να εξερευνήσει το χώρο αυτόν προς αναζήτηση αυτής της πιθανής καλύτερης ενέργειας. Το ποια είναι η σωστή ισορροπία έχει μελετηθεί, μελετάται εκτενώς και σαφώς εξαρτάται από το πρόβλημα αλλά κυρίως είναι κάτι που προκύπτει από την εμπειρία του παίκτη λόγω επιλύσεων αρκετών άλλων προβλημάτων. Η ισορροπία αυτή όμως δεν είναι εφικτή με όλες τις μεθόδους μάθησης, όπως στις μεθόδους επιβλεπόμενης μάθησης.

Για να γίνει πιο κατανοητό αυτό το ζήτημα της ισορροπίας, ας επανέλθουμε στο παράδειγμα του Tetris. Έστω ότι ο παίκτης βρίσκεται σε κάποια κατάσταση και έχει κάποιες διαθέσιμες επιλογές για να προχωρήσει: να ψάξει ανάμεσα στις διαθέσιμες κινήσεις που έχει για την πιο επικερδή όσον αφορά την ανταμοιβή ή να εξερευνήσει τον χώρο αναζήτησης περαιτέρω προχωρώντας σε επόμενες καταστάσεις του συστήματος τοποθετώντας άλλα τούβλα σε καινούριες θέσεις ή να συνδυάσει τα δυο προηγούμενα βάσει κάποιου κανόνα. Για να επιτευχθεί η επιθυμητή ισορροπία, θα πρέπει να επιλέξει την τρίτη επιλογή χρησιμοποιώντας έναν εμπειρικό κανόνα ή τον κανόνα που του έχει δοθεί από τον επιβλέποντα.

Ένα άλλο βασικό χαρακτηριστικό της Ενισχυτικής Μάθησης είναι ότι οι μέθοδοί της έχουν μια σφαιρική οπτική των προβλημάτων κατά τα οποία ο παίκτης κατευθύνεται προς έναν στόχο υπό αβέβαιο περιβάλλον διεπαφής. Αυτό έρχεται σε αντίθεση με τις άλλες μεθόδους εκμάθησης οι οποίες επικεντρώνονται στην επίλυση συγκεκριμένων υποπροβλημάτων χωρίς να δίνουν σημασία στο πως αυτές οι επιλύσεις μπορούν να γενικευθούν .

2.2 Συστατικά Ενισχυτικής Μάθησης

Τα συστατικά της Ενισχυτικής Μάθησης είναι τα εξής : ο παίκτης, το περιβάλλον διεπαφής του, η πολιτική του, η συνάρτηση ανταμοιβής, η συνάρτηση αξιολόγησης και εάν δίνεται, το μοντέλο του περιβάλλοντος διεπαφής. Αναλυτικότερα :

Παίκτης

Όλοι οι παίκτες προβλημάτων Ενισχυτικής Μάθησης έχουν συγκεκριμένους στόχους, μπορούν να αντιληφθούν έως έναν βαθμό κάποια στοιχεία του περιβάλλοντός τους και μπορούν να επιλέγουν ενέργειες για να το επηρεάσουν μέσω της προσαρμογής τους στις αλλαγές του. Επιπρόσθετα, συνήθως ισχύει η υπόθεση ότι οι παίκτες πρέπει να δρουν υπό συνθήκες σημαντικής αβεβαιότητας όσον αφορά το περιβάλλον τους. Όταν απαιτείται προγραμματισμός, ο παίκτης πρέπει να χειριστεί την σχέση μεταξύ προγραμματισμού και επιλογής ενεργειών σε πραγματικό χρόνο καθώς και του ερωτήματος του πώς τα μοντέλα του περιβάλλοντος αποκτώνται και βελτιώνονται. Στο συγκεκριμένο πρόβλημά μας, ο παίκτης είναι το πρόγραμμα του παιχνιδιού Tetris.

Περιβάλλον διεπαφής

Το περιβάλλον διεπαφής είναι στην ουσία ότι δεν μπορεί να επηρεάσει ο παίκτης. Αυτό είναι μια σημαντική παρατήρηση η οποία καθορίζει την σωστότερη διατύπωση και συνεπώς επίλυση του κάθε προβλήματος καθώς και τον σωστότερο καθορισμό των ορίων μεταξύ παίκτη και περιβάλλοντος διεπαφής. Το περιβάλλον μπορεί να είναι στατικό (ανεξάρτητο από τον χρόνο) ή δυναμικό (εξαρτημένο από τον χρόνο). Ανάλογα με τους κανόνες που το διέπουν, το στοιχείο αυτό είναι που καθορίζει τις επιλογές και τις καταστάσεις του παίκτη κατά την διάρκεια της διαδικασίας επίτευξης του στόχου του και σε έναν βαθμό τον χρόνο που έχει διαθέσιμο. Το περιβάλλον είναι επίσης αυτό που δίδει στον παίκτη την επόμενη κατάσταση και ανταμοιβή μετά από μια ενέργεια του. Στο Tetris, το περιβάλλον αυτό είναι ο πίνακας στον οποίο βλέπουμε να διαδραματίζεται το παιχνίδι, δηλαδή ο πίνακας όπου τα τετρόνιμα πέφτουν ένα ένα με τη σειρά.

Πολιτική

Η πολιτική του παίκτη είναι η συμπεριφορά του σε κάθε χρονική στιγμή και κατάσταση. Με άλλα λόγια, η πολιτική είναι η αντιστοίχιση καταστάσεων του συστήματος με συγκεκριμένες ενέργειες που πρέπει να γίνουν όταν ο παίκτης βρεθεί στις καταστάσεις αυτές. Η πολιτική αυτή μπορεί να είναι απλές συναρτήσεις ή πίνακες με συγκεκριμένες τιμές ή να περιλαμβάνουν πολύπλοκους υπολογισμούς όπως σε μια διαδικασία αναζήτησης. Η πολιτική αποτελεί τον πυρήνα της Ενισχυτικής Μάθησης με την έννοια ότι από μόνη της επαρκεί για τον καθορισμό της συμπεριφοράς του παίκτη. Ως επί το πλείστον, η πολιτική που ακολουθείται είναι στοχαστική. Είναι σαφές από τα παραπάνω ότι λόγω της αποκτούμενης μετά από ένα χρονικό διάστημα

εμπειρίας του παίκτη η πολιτική του θα αλλάξει όταν αυτός κρίνει ότι με την αλλαγή αυτή θα φτάσει ταχύτερα στον στόχο του. Στο Tetris η πολιτική του παίκτη είναι να εφαρμόσει τις κατάλληλες κινήσεις την κατάλληλη στιγμή ώστε να αποκομίσει την μεγαλύτερη δυνατή ανταμοιβή. Η ενέργεια του παίκτη εξαρτάται κάθε φορά από το ποιες τιμές έχουν λάβει οι παράγοντες που επηρεάζουν την πολιτική αυτή που ονομάζονται χαρακτηριστικές συναρτήσεις του παιχνιδιού Tetris.

Συνάρτηση αθροιστικών ανταμοιβών

Η συνάρτηση αθροιστικών ανταμοιβών καθορίζει τον στόχο σε ένα πρόβλημα Ενισχυτικής Μάθησης. Συγκεκριμένα, αντιστοιχεί κάθε κατάσταση (ή ζεύγος κατάστασης-ενέργειας) του περιβάλλοντος με έναν απλό αριθμό, μια ανταμοιβή, υποδεικνύοντας το πόσο επιθυμητή είναι αυτή. Ο μοναδικός στόχος του παίκτη είναι να μεγιστοποιήσει το συνολικό άθροισμα των ανταμοιβών που δέχεται μακροπρόθεσμα. Με την συνάρτηση αυτή οι διαθέσιμες ενέργειες διαχωρίζονται σε καλές και κακές. Όπως γίνεται αντιληπτό, αποτελούν τα άμεσα και καθοριστικά χαρακτηριστικά του περιβάλλοντος. Για αυτό τον λόγο δεν πρέπει να μπορεί ο παίκτης να αλλάζει τη συνάρτηση αυτή που αποτελεί όως τη βάση για την αλλαγή της πολιτικής. Επομένως, εάν μια ενέργεια συνοδεύεται από μικρή ανταμοιβή, τότε ο παίκτης θα προτιμήσει μια άλλη ενέργεια με υψηλότερη ανταμοιβή αλλάζοντας την πολιτική του. Γενικώς, οι συναρτήσεις ανταμοιβών είναι στοχαστικές. Άρα, με βάση όσα γράφησαν για την συνάρτηση αθροιστικών ανταμοιβών, αυτή υποδεικνύει το τι είναι καλύτερο για τον παίκτη βραχυπρόθεσμα. Μια μαθηματική έκφραση της συνάρτησης αυτής σύμφωνα με τα παραπάνω είναι :

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + ... + r_T$$

Και για την περίπτωση των αποσβένουσων αθροιστικών ανταμοιβών :

$$R_t = r_{t+1} + \gamma * r_{t+2} + \gamma^2 * r_{t+3} + ... + \gamma^{T-t-1} * r_T$$

ή

$$R_t = \sum_{k=0}^{T} \gamma^k * r_{t+k+1}$$

Η συνάρτηση ανταμοιβής στο συγκεκριμένο μας πρόβλημα λειτουργεί ως εξής : αρχικά, δίδεται για ένα σημαντικό αριθμό αρχικών χρονικών βημάτων μηδενική ανταμοιβή ώστε ο παίκτης να αναγκαστεί να μάθει όσο το δυνατόν πιο γρήγορα και καλά προκειμένου να αποκτήσει αργότερα ανταμοιβή. Έπειτα, μετά από ένα αριθμό χρονικών βημάτων που καθορίζεται από μια συνάρτηση πιθανότητας ως προς τον αριθμό αυτό, δίδεται μια ανταμοιβή ώστε να συνεχίσει την ”καλή” πορεία του ο παίκτης. Επίσης, υπάρχει ένας κύκλος που ακολουθείται για τη εξαγωγή της ανταμοιβής, δηλαδή ενώ εξελίσσεται το παιχνίδι παρατηρείται μια αυξανόμενη σειρά απόδοσης της ανταμοιβής και όταν φθάσει τη μέγιστη τιμή τότε ξαναγυρνά η συνάρτηση στην μηδενική ανταμοιβή. Η λογική είναι να μην συνεχίζει να εκμεταλλεύεται μια «καλή» πορεία μόνο ο παίκτης αλλά με τον μηδενισμό της ανταμοιβής ανά τακτά χρονικά διαστήματα να αναγκάζεται να εξερευνήσει άλλες πιθανές πορείες. Τέλος, στα τελευταία χρονικά βήματα

δίδονται μηδενικές ανταμοιβές διότι θεωρείται ότι ο παίκτης με το παίξιμό του έχει φτάσει σε σημείο κορεσμού όσον αφορά την πορεία του και την ανταμοιβή που έχει πάρει.Όλες οι ανταμοιβές λαμβάνονται βάσει ενός διανύσματος που είναι το εξής:[0, 0.441697, 1.3251, 2.65018, 3.975265]

Συνάρτηση αξιολογήσεων

Η συνάρτηση αξιολογήσεων σε αντίθεση με την συνάρτηση ανταμοιβών, υποδεικνύει το τι είναι καλύτερο με την μακροπρόθεσμη έννοια. Με άλλα λόγια, η αξιολόγηση μιας κατάστασης είναι το αναμενόμενο άθροισμα των ανταμοιβών που θα λάβει ο παίκτης εάν ξεκινήσει από την κατάσταση αυτή και προχωρήσει ακολουθώντας την πολιτική που έχει. Για παράδειγμα, μια κατάσταση μπορεί να έχει μια χαμηλή άμεση ανταμοιβή άλλα οι ανταμοιβές που θα επακολουθήσουν να είναι υψηλές οπότε συνολικά μακροπρόθεσμα να είναι επιθυμητή η κατάσταση αυτή.

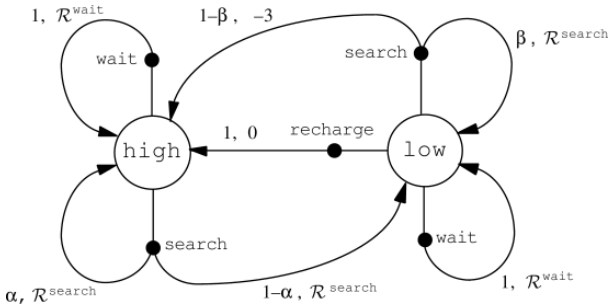
Μια σημαντική παρατήρηση είναι ότι οι ανταμοιβές έχουν πρωταρχική σημασία ενώ οι αξίες δευτερεύουσα και ο λόγος είναι ότι χωρίς τις ανταμοιβές δεν έχει έννοια η αξιολόγηση εφόσον η αξιολόγηση αποσκοπεί στην μεγιστοποίηση μελλοντικών ανταμοιβών. Παρόλα αυτά, η παρούσα εργασία ασχολείται με τις αξιολογήσεις κατά την λήψη και αξιολόγηση αποφάσεων διότι είναι προβλέψεις της επιθυμητότητας των καταστάσεων μέσω του συνολικού τους αθροίσματος ανταμοιβών. Οι ενέργειες επιλέγονται βάσει των αξιών και όχι των ανταμοιβών τους. Όμως, ενώ οι ανταμοιβές δίδονται εξαρχής από το περιβάλλον, οι υπολογισμοί των αξιών γίνονται με εκτιμήσεις που επαναλαμβάνονται αρκετές φορές επομένως είναι πολύ πιο δύσκολος ο υπολογισμός τους. Μάλιστα το πιο σημαντικό τμήμα των αλγορίθμων Ενισχυτικής Μάθησης είναι οι μέθοδοι εκτίμησης των αξιών. Η συνάρτηση αξιών του Tetris είναι ένα άθροισμα γινομένων μεταξύ των βαρών που αντιστοιχούν στις χαρακτηριστικές συναρτήσεις του παιχνιδιού και των συναρτήσεων αυτών και το αριθμητικό αποτέλεσμα της συνάρτησης αξιών είναι η ποσοτική εκτίμηση της κατάστασης στην οποία βρίσκεται ο παίκτης.

Μοντέλο περιβάλλοντος

Το μοντέλο του περιβάλλοντος διεπαφής με τον παίκτη είναι στην ουσία μια μίμηση της συμπεριφοράς του περιβάλλοντος. Δηλαδή, δίδοντας στο μοντέλο συγκεκριμένη κατάσταση και ενέργεια αυτό προβλέπει την επόμενη κατάσταση και ανταμοιβή. Τα μοντέλα χρησιμοποιούνται για προγραμματισμό, δηλαδή για την διαδικασία επιλογής μιας ενέργειας λαμβάνοντας υπόψη μελλοντικές καταστάσεις πριν καν αυτές εμφανιστούν.

2.3 Μαρκοβιανές διαδικασίες αποφάσεων

Οι Μαρκοβιανές διαδικασίες αποφάσεων είναι ένα μοντέλο εφαρμογής Ενισχυτικής Μάθησης που ικανοποιεί την Μαρκοβιανή ιδιότητα. Η Μαρκοβιανή ιδιότητα είναι η ιδιότητα που έχει μια ανταμοιβή να περιλαμβάνει όλα τα προηγούμενα σήματα στην τωρινή κατάσταση του



Σχήμα 2.1: Μια Μαρκοβιανή διαδικασία.

συστήματος χωρίς να χάσει καμία πληροφορία αλλά η τωρινή αυτή κατάσταση να εξαρτάται μόνο από την προηγούμενη της και όχι από όλες τις προηγούμενες της. Όταν οι χώροι των ενεργειών και των καταστάσεων είναι πεπερασμένοι, τότε ο στόχος Ενισχυτικής Μάθησης ονομάζεται πεπερασμένη Μαρκοβιανή διαδικασία απόφασης. Οι διαδικασίες αυτές έχουν πολύ μεγάλη σημασία για την Ενισχυτική Μάθηση και για αυτό μελετώνται εκτενώς και καθορίζονται από το πώς είναι κατανεμημένες οι καταστάσεις και οι ενέργειες καθώς και από τις δυναμικές του περιβάλλοντος σε διάστημα ενός χρονικού βήματος.

Για δεδομένη κατάσταση και ενέργεια, s και a, η πιθανότητα κάθε άλλης πιθανής κατάστασης είναι :

$$P^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$$

Οι πιθανότητες αυτές λέγονται πιθανότητες μετάβασης του συστήματος και αποτελούν το προαναφερθέν **μοντέλο** του συστήματος.

Ομοίως, η μαθηματική έκφραση της αναμενόμενης αξιολόγησης της μελλοντικής ανταμοιβής είναι :

$$R^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}$$

Για τις Μαρκοβιανές διαδικασίες αποφάσεων χρησιμοποιούνται δυο επαναληπτικές εξισώσεις :

$$V^\pi(s) = E_\pi\{R_t | s^t = s\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k * r_{t+k+1} | s_t = s\}$$

όπου E_π είναι η αναμενόμενη αξιολόγηση δεδομένου ότι ο παίκτης ακολουθεί την πολιτική π. Ονομάζεται η συνάρτηση $V^\pi(s)$ συνάρτηση κατάστασης-αξιολόγησης για την πολιτική π, τη συνάρτηση αξιών του συστήματος δηλαδή.

Η δεύτερη εξίσωση είναι :

$$Q^\pi(s,a) = E_\pi\{R_t | s^t = s, a_t = a\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k * r_{t+k+1} | s_t = s, a_t = a\}$$

όπου $Q^{\pi}(s,a)$ είναι η συνάρτηση ενέργειας-αξιολόγησης για την πολιτική π, εκτιμά δηλαδή το πόσο καλή είναι μια συγκεκριμένη ενέργεια εάν πραγματοποιηθεί. Συνήθως οι εκτιμήσεις αυτές γίνονται εμπειρικά μετά από αρκετές επαναλήψεις, όπως για παράδειγμα αφού συλλεχτούν επαρκή αθροίσματα ανταμοιβών μετά λαμβάνεται ως εκτίμηση ο μέσος όρος τους. Η βασική διαφορά μεταξύ των δυο αυτών συναρτήσεων εκτίμησης είναι ότι η συνάρτηση $V^{\pi}(s)$ είναι εκτιμήτρια της κατάστασης s υπό πολιτική π ενώ η συνάρτηση $Q^{\pi}(s,a)$ είναι η εκτιμήτρια της ενέργειας a από την κατάσταση s_t στην κατάσταση s_{t+1} και μετά ακολουθείται η πολιτική π .

Οι παραπάνω συναρτήσεις εκτιμώνται από τις εξισώσεις Bellman οι οποίες είναι οι εξής :

$$V^{\pi}(s) = \sum_{a} \pi_{s,a} * \sum_{s'} P^{a}_{ss'} * (R^{a}_{ss'} + \gamma * V^{\pi}(s'))$$

που είναι η επαναληπτική εξίσωση για την εκτίμηση της συνάρτησης αξιολόγησης μιας κατάστασης. Εκφράζει την σχέση μεταξύ της αξιολόγησης μιας τωρινής κατάστασης και των αξιών των καταστάσεων που εμφανίζονται μετά από αυτήν. Πιο συγκεκριμένα, η εξίσωση Bellman αθροίζει όλες τις πιθανές περιπτώσεις των ακολουθιών των ενεργειών που μπορεί να εκτελέσει ο παίκτης και τις πολλαπλασιάζει με την πιθανότητα εμφάνισής τους. Έτσι, δηλώνει ότι η αξιολόγηση της αρχικής κατάστασης πρέπει να είναι ίση με το άθροισμα της αναμενόμενης αλλά μειωμένης λόγω απόσβεσης αξιολόγησης της επόμενης κατάστασης και της ανταμοιβής που αναμένεται να αποδοθεί κατά την διαδρομή αυτή. Επίσης, η συνάρτηση αξιολόγησης $V^{\pi}(s)$ αποτελεί μοναδική λύση για την εξίσωση Bellman. Η δεύτερη εξίσωση είναι η εκτιμήτρια για την συνάρτηση αξιολόγησης-ενέργειας και είναι :

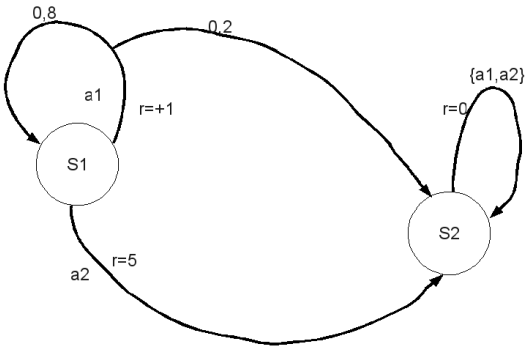
$$Q^{\pi}(s,a) = \sum_{a} \pi_{s,a} * \sum_{s'} P^{a}_{ss'} * (R^{a}_{ss'} + \gamma * Q^{\pi}(s',a'))$$

όπου πάλι εδώ υπάρχει παρόμοια με πριν αναδρομική σχέση για τα ζεύγη των αξιών-ενεργειών.Επίσης υπάρχει στην εξίσωση ένα επιπρόσθετο άθροισμα εφόσον η αναφορά γίνεται σε ζεύγη αξιών-ενεργειων και όχι μόνο σε αξίες.

Για να βρεθεί η βέλτιστη πολιτική, πρέπει να βρεθεί η συνάρτηση αξιολογήσεων η οποία να επιφέρει το μεγαλύτερο άθροισμα ανταμοιβών, επομένως και η πολιτική που θα ακολουθείται για τον υπολογισμό της συνάρτησης αυτής θα είναι και η βέλτιστη. ΄Έτσι, οι εξισώσεις Bellman θα μετασχηματιστούν ως εξής για την εύρεση της βέλτιστης πολιτικής :

$$V^{\pi *}(s) = \max \sum_{s'} P^{a}_{ss'} * (R^{a}_{ss'} + \gamma * V^{\pi *}(s'))$$

$$Q^{\pi *}(s,a) = \sum_{s'} P^{a}_{ss'} * (R^{a}_{ss'} + \gamma * \max_{a'} Q^{\pi *}(s',a'))$$



Σχήμα 2.2: Γράφημα μεταβάσεων μεταξύ των δύο καταστάσεων του παραδείγματός μας .

2.4 Παράδειγμα κατανόησης των προηγούμενων ενοτήτων

Στην ενότητα αυτή θα παρουσιαστεί ένα παράδειγμα για να γίνουν κατανοητές οι έννοιες και εξισώσεις που προαναφέρθηκαν. Καταρχάς, παρουσιάζεται το πρόβλημα σε γραφική μορφή (βλέπε Σχήμα 2.1):

Αυτό είναι το γράφημα μεταβάσεων του παραδείγματος, δηλαδή είναι η γραφική απεικόνιση των καταστάσεων και των πιθανών ενεργειών που οδηγούν από την μια στην άλλη. Εφόσον δίδονται:

- το γράφημα μεταβάσεων,
- υπάρχουν μεταβάσεις που να οδηγούν από την μία κατάσταση στην άλλη,
- και η αξιολόγηση μιας τωρινής κατάστασης εξαρτάται μόνο από την προηγούμενή της, το πρόβλημά μας είναι μια Μαρκοβιανή διαδικασία απόφασης. Επίσης, επειδή δίνονται και οι πιθανότητες μετάβασης μεταξύ των καταστάσεων είναι γνωστό και το **μοντέλο** του προβλήματος. Συγκεκριμένα, αυτό είναι :

$$p(s1|s1,a1) = 0.8, p(s2|s1,a1) = 0.2, p(s1|s1,a2) = 0,$$
$$p(s2|s1,a2) = 1, p(s2|s2,a1) = 1, p(s2|s2,a2) = 1$$

. Επίσης, είναι γνωστές και οι ανταμοιβές που δίδει το περιβάλλον στον παίκτη όταν αυτός εκτελέσει κάποια ενέργεια και αυτές είναι :

$$R(s1,a1) = 1, R(s1,a2) = 5, R(s2,\bigstar) = 0$$

όπου ★ είναι οποιαδήποτε ενέργεια.

Επιπλέον, το σύνολο καταστάσεων είναι το S=s1,s2 και το σύνολο ενεργειών A=a1,a2. Το ζητούμενο είναι να βρεθεί η βέλτιστη πολιτική, δηλαδή οι αντιστοιχίες ενεργειών και καταστάσεων που θα επιφέρουν το μεγαλύτερο άθροισμα ανταμοιβών. Επιπρόσθετα, στο πρόβλημα μας όταν ο παίκτης βρεθεί στην κατάσταση s2 τότε δεν μπορεί να ξεφύγει από αυτή. Η s2 ονομάζεται απορροφητική κατάσταση. Έτσι, έχει νόημα μόνο ο υπολογισμός της βέλτιστης πολιτικής ξεκινώντας από την κατάσταση s1. Παρακάτω θα αναλυθεί ο υπολογισμός της βέλτιστης πολιτικής στην αντίστοιχη ενότητα.

2.5 Μέθοδοι επίλυσης

Στην ενότητα αυτή παρουσιάζονται κάποιες μέθοδοι επίλυσης προβλημάτων Ενισχυτικής μάθησης μέσω Μαρκοβιανών διαδικασιών απόφασης.

Επαναληπτική αξιολόγηση πολιτικής

Καταρχάς, ας υπενθυμίσουμε πως υπολογίζεται η συνάρτηση αξιολογήσεων:

$$V^{\pi}(s) = \sum_{a} \pi_{s,a} * \sum_{s'} P^{a}_{ss'} * (R^{a}_{ss'} + \gamma * V^{\pi}(s'))$$

Η εξίσωση αυτή χρησιμοποιείται ώστε να αξιολογηθεί η πολιτική που ακολουθείται.Εάν το μοντέλο του περιβάλλοντος είναι γνωστό τότε η παραπάνω εξίσωση είναι ένα σύστημα γραμμικών εξισώσεων που μπορεί να επιλυθεί. Ο αριθμός των εξισώσεων είναι ίσος με τον αριθμό των καταστάσεων του προβλήματος. Πιο κατάλληλες για τις περιπτώσεις που μελετάμε είναι οι μέθοδοι επαναληπτικής επίλυσης. Μια πιο ολοκληρωμένη μορφή της εξίσωσης αυτής ώστε να φαίνεται ο επαναληπτικός χαρακτήρας των μεθόδων είναι :

$$V_{k+1}(s) = \sum_{a} \pi_{s,a} * \sum_{s'} P^{a}_{ss'} * (R^{a}_{ss'} + \gamma * V_{k}(s'))$$

Προφανώς, πρέπει $V_k = V^{\pi}$να είναι ένα σταθερό σημείο για αυτήν την επαναληπτική εξίσωση ώστε να υπολογιστεί το V_{k+1} και η σύγκλιση στο V^{π} είναι εγγυημένη όταν $k \rightarrow \equiv \infty$. Η διαδικασία αυτή ονομάζεται **επαναληπτική αξιολόγηση πολιτικής**.

Βελτίωση πολιτικής

Η Βελτίωση πολιτικής χρησιμοποιείται διότι δίνει λύσεις στο εξής δίλημμα : Όταν έχουμε μια πολιτική π, αυτή μας υποδεικνύει ποιες ενέργειες να κάνουμε όταν βρισκόμαστε σε κάθε κατάσταση. Οι ενέργειες αυτές όμως είναι και οι βέλτιστες ή μήπως υπάρχουν άλλες ενέργειες για τις συγκεκριμένες καταστάσεις που είναι καλύτερες; Επομένως, πρέπει να εισαχθεί στην μοντελοποίηση μας ο εξής περιορισμός :

$$Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s)$$

. Οπότε πρέπει η πολιτική π' να είναι το ίδιο ή καλύτερη από την πολιτική π. Ο περιορισμός αυτός ισοδυναμεί και με τον εξής περιορισμό :

$$V^{\pi'}(s) \geq V^{\pi}(s)$$

Οι περιορισμοί αυτοί αποτελούν και το θεώρημα βελτίωσης της πολιτικής. Επηρεασμένος από τις συνθήκες αυτές, είναι φυσικό να οδηγηθεί κανείς στο να σκεφτεί να βελτιώσει όλες τις πιθανές ενέργειες για κάθε κατάσταση, πάντα ως προς το $Q^{\pi}(s, a)$ και άρα να μετατρέψει την αρχική πολιτική π σε μια άπληστη πολιτική π'. Οπότε η νέα πολιτική π' υπολογίζεται ως εξής:

$$\pi'(s) = arg\max_{a} Q^{\pi}(s, a) = arg\max_{a} \sum_{s'} P^{a}_{ss'} * (R^{a}_{ss'} + \gamma * V^{\pi}(s'))$$

Επανάληψη πολιτικής

Η όλη ιδέα αυτής της μεθόδου είναι η εξής: Έστω ότι έχουμε μια πολιτική π. Την αξιολογούμε με την γνωστή μας συνάρτηση $V^{\pi}(s)$ και έπειτα εφαρμόζουμε την Βελτίωση πολιτικής, δηλαδή κάνουμε την πολιτική μας άπληστη, την αξιολογούμε και πάλι και η διαδικασία αυτή επαναλαμβάνεται.

Μέθοδος Μάθησης συνάρτησης ζεύγους αξιολόγησης-ενέργειας

Η μέθοδος αυτή περιγράφεται από την παρακάτω αναδρομική εξίσωση:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + a[r_{t+1} + \gamma * \max_a Q(s_{t+1}, a) - Q(s_t, a_t)].$$

. Με αυτόν τον τρόπο, η συνάρτηση ζεύγους αξιολόγησης-ενέργειας προσεγγίζει απευθείας το Q^*, την βέλτιστη συνάρτηση ζεύγους αξιολόγησης-ενέργειας, ανεξαρτητα από την πολιτική που ακολουθείται.

Sarsa μέθοδος

Η μέθοδος Sarsa είναι μια μέθοδος εκτίμησης της συνάρτησης ζεύγους αξιολόγησης-ενέργειας και συνίσταται στην εκτίμηση της τωρινής πολιτικής μέσω της συνάρτησης ζεύγους αξιολόγησης-ενέργειας και στην ταυτόχρονη μετατροπή της πολιτικής αυτής σε άπληστη πολιτική όσον αφορά την συνάρτηση αυτή. Η μέθοδος περιγράφεται από την εξής αναδρομική εξίσωση :

$$Q^{\pi}(s_t, a_t) \leftarrow Q^{\pi}(s_t, a_t) + \alpha * [r_{t+1} + \gamma * Q^{\pi}(s_{t+1}, a_{t+1}) - Q^{\pi}(s_t, a_t)]$$

Η μέθοδος ονομάστηκε Sarsa διότι κάθε φορά το σύνολο $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ ανανεώνεται.

Υπολογισμός βέλτιστης πολιτικής του παραδείγματος

Στην υποενότητα αυτήν θα παρουσιαστούν υλοποιήσεις των μεθόδων επίλυσης: Επαναληπτική αξιολόγηση πολιτικής, Βελτίωση πολιτικής και Επανάληψη πολιτικής. Η Επαναληπτική αξιολόγηση πολιτικής είναι στην ουσία ο επαναληπτικός υπολογισμός της συνάρτησης αξιολόγησης μιας κατάστασης. Η υπόθεση είναι ότι η πολιτική του παίκτη όταν βρίσκεται στην κατάσταση s1 να εκτελεί την ενέργεια a1, ο συντελεστής απόσβεσης γ=1 και ε=0.1, όπου ε η διαφορά μεταξύ δύο διαδοχικών υπολογισμών της συνάρτησης αξιολόγησης της κατάστασης s1: Το $V(s2) = 0$ διότι είναι μια απορροφητική κατάσταση και άρα δίδει μηδενική ανταμοιβή.

$$V_1(s1) = P^{a1}_{s1s1} * (R^{a1}_{s1s1} + V_0(s1)) + P^{a1}_{s1s2} * (R^{a1}_{s1s2} + V_0(s2)) = 0.8 * (1+0) + 0.2 * (1+0) = 1$$

$$V_2(s1) = 0.8 * (1+1) + 0.2 * (1+0) = 1.8$$

$$V_3(s1) = 0.8 * (1 + 1.8) + 0.2 * (1 + 0) = 2.44$$

$$V_4(s1) = 0.8 * (1 + 2.44) + 0.2 * (1 + 0) = 2.952$$

$$V_5(s1) = 0.8 * (1 + 2.952) + 0.2 * (1 + 0) = 3.362$$

$$V_6(s1) = 0.8 * (1 + 3.362) + 0.2 * (1 + 0) = 3.689$$

$$V_7(s1) = 0.8 * (1 + 3.689) + 0.2 * (1 + 0) = 3.951.....$$

Για $k \to \infty$ το $V^{\pi}(s1)$ είναι σίγουρα > 5.

Προκειμένου να διαπιστωθεί εάν η πολιτική αυτή είναι η βέλτιστη, θα εφαρμοστεί η Βελτίωση Πολιτικής για να δούμε εάν η πολιτική που ακολουθήθηκε επιδέχεται βελτίωση. Συγκεκριμένα, έστω ότι μια νέα πολιτική είναι ο παίκτης να εκτελεί την ενέργεια a2 όταν βρίσκεται στην κατάσταση s1 :

$$Q^{\pi'}(s1, a2) = P^{a2}_{s1s2} * (R^{a2}_{s1s2} + V^{\pi'}(s2)) = 1 * (5 + 0) = 5$$

. Όμως, το

$$Q^{\pi'}(s1, a2) < V^{\pi}(s1)$$

άρα είναι καλύτερη πολιτική η $\pi(s1)$.

Επομένως καταλήγουμε ότι ο παίκτης οταν βρεθεί στην κατάσταση s1 πρέπει να εκτελεί την ενέργεια a1 ώστε να συλλέξει περισσότερη ανταμοιβή μέχρι να καταλήξει στην απορροφητική κατάσταση s2, όπου προφανώς θα τερματιστεί και η διαδικασία.

Κεφάλαιο 3

Γενίκευση και προσέγγιση συναρτήσεων

3.1 Ορισμός γενίκευσης και πρόβλεψης αξιών μέσω των προσεγγίσεων συναρτήσεων

Η λογική και η διαδικασία κατασκευής της συνάρτησης αξιών των καταστάσεων του προβλήματος υπό μελέτη είναι ακριβώς η ίδια όπως και σε όλες τις προηγούμενες μεθόδους. Υπάρχουν προβλήματα στα οποία:

* η συνάρτηση αξιολόγησης μπορεί να υπολογίζεται βάσει ενός δένδρου αποφάσεων και το $\overrightarrow{\theta_t}$ να αποτελείται από όλες τις παραμέτρους που καθορίζουν τα σημεία διαίρεσης των κλαδιών και τις αξίες των τελικών κόμβων των δένδρων ή φύλλων όπως ονομάζονται.
* η συνάρτηση αξιολόγησης μπορεί να υπολογίζεται από ένα τεχνητό νευρωνικό δίκτυο με $\overrightarrow{\theta_t}$ να είναι το διάνυσμα των βαρών σύνδεσης των νευρώνων. Η μόνη διαφορά είναι ότι η προσεγγιστική συνάρτηση αξιολόγησης V_t δεν ανήκει στις παραπάνω περιπτώσεις αλλά είναι μια παραμετρική συναρτησιακή μορφή με διάνυσμα παραμέτρων θ_t. Με άλλα λόγια, η συνάρτηση αξιολόγησης εξαρτάται πλήρως από το $\overrightarrow{\theta_t}$ και μεταβάλλεται σε κάθε χρονικό βήμα μόνο όπως μεταβάλλεται αυτό. Τυπικά, ο αριθμός των παραμέτρων (των στοιχείων του διανύσματος $\overrightarrow{\theta_t}$ δηλαδή) είναι αρκετά μικρότερος του αριθμού των καταστάσεων και με την ανανέωση μιας παραμέτρου αλλάζει η εκτιμώμενη αξιολόγηση πολλών καταστάσεων. Αυτό γίνεται διότι με αυτόν τον τρόπο όταν η αξιολόγηση μιας κατάστασης ανανεώνεται, η αλλαγή αυτή γενικεύεται από την δεδομένη κατάσταση και στις άλλες καταστάσεις(**Αυτή είναι η έννοια της γενίκευσης.**). Είναι φυσικό η ανανέωση αυτή να θεωρηθεί ως παράδειγμα της επιθυμητής συμπεριφοράς εισόδου-εξόδου δεδομένων της εκτιμώμενης συνάρτησης αξιολόγησης. Έως τώρα, η διαδικασία της ανανέωσης ήταν τετριμμένη: η αξιολόγηση του στοιχείου του πίνακα της συνάρτησης αξιών που αντιστοιχεί στην συγκεκριμένη κατάσταση απλώς αυξανόταν κατά ένα ποσοστό της διαφοράς της από την αξιολόγηση της ”ανανεωμένης τιμής” που είναι και η τιμή-

στόχος. Πλέον, η ανανέωση γίνεται με την χρήση αυθαίρετα πολύπλοκων και εξελιγμένων μεθόδων προσέγγισης συναρτήσεων. Συγκεκριμένα, ως δεδομένα εισόδου σε αυτές τις συναρτήσεις χρησιμοποιούνται παραδείγματα της επιθυμητής συμπεριφοράς εισόδου-εξόδου της εκτιμώμενης συνάρτησης αξιολόγησης που προσπαθούν να προσεγγίσουν. Επίσης, τις χρησιμοποιούμε για την πρόβλεψη αξιών απλώς με το να τις εκπαιδεύουμε με τα παραδείγματα αυτά.
Έπειτα, θεωρούμε την προσεγγιστική συνάρτηση που παράγουν ως μια εκτιμώμενη συνάρτηση αξιολόγησης. Μια σημαντική παρατήρηση είναι ότι ναι μεν με τη θεώρηση των ανανεώσεων ως παραδειγμάτων εκπαίδευσης μπορούμε να χρησιμοποιήσουμε μεθόδους προσέγγισης από ένα ευρύ φάσμα υπάρχοντων μεθόδων για την πρόβλεψη αξιών αλλά από την άλλη δε δεν είναι όλες τους το ίδιο ικανοποιητικές. Ο λόγος είναι ότι στην Ενισχυτική Μάθηση είναι σημαντικό να λαμβάνει χώρα η μάθηση κατά την διάρκεια της αλληλεπίδρασης του παίκτη με το περιβάλλον του ή με το μοντέλο του περιβάλλοντός του. Επομένως, απαιτούνται μέθοδοι που να μπορούν να μαθαίνουν από αυξανόμενα σε αριθμό δεδομένα εισόδου. Επιπρόσθετα, η Ενισχυτική Μάθηση απαιτεί από τις μεθόδους αυτές και να μπορούν να χειριστούν μη στατικές συναρτήσεις στόχων (συναρτήσεις στόχων δηλαδή που αλλάζουν με τον χρόνο).

Μέτρο απόδοσης των μεθόδων προσέγγισης συναρτήσεων

Το μέτρο για την ορθή αξιολόγηση των μεθόδων προσέγγισης συναρτήσεων είναι η ελαχιστοποίηση του σφάλματος ελαχίστων τετραγώνων μιας κατανομής P των δεδομένων εισόδου. Στο πρόβλημα της πρόβλεψης αξιών της περίπτωσής μας, τα δεδομένα εισόδου είναι οι καταστάσεις, η συνάρτηση-στόχος είναι η πραγματική συνάρτηση αξιών V^{π} και το σφάλμα ελαχίστων τετραγώνων για μια προσέγγιση V_t με διάνυσμα παραμέτρων $\overrightarrow{\theta_t}$ είναι

$$MSE(\overrightarrow{\theta_t}) = \sum_{s \in S} P(s) * [V^{\pi}(s) - V_t(s)]^2$$

όπου P είναι μια κατανομή για την στάθμιση των σφαλμάτων των διαφόρων καταστάσεων. Αυτή η κατανομή είναι σημαντική διότι συνήθως δεν είναι εφικτό να μηδενιστεί το σφάλμα σε όλες τις καταστάσεις. Ούτως ή άλλως, οι καταστάσεις είναι πολύ περισσότερες από τα στοιχεία του διανύσματος $\overrightarrow{\theta_t}$. Γενικώς, μπορεί να επιτευχθεί καλύτερη προσέγγιση σε κάποιες καταστάσεις σε βάρος άλλων καταστάσεων οι οποίες έχουν χειρότερη προσέγγιση. Αυτό το δούναι και λαβείν καθορίζεται από την κατανομή P με την οποία επιλέγονται οι καταστάσεις των εκπαιδευτικών παραδειγμάτων και είναι ίδια με τη κατανομή των καταστάσεων των οποίων οι αξίες ανανεώνονται. Επίσης, εάν είναι επιθυμητή η ελαχιστοποίηση σφάλματος για ορισμένη κατανομή καταστάσεων, τότε είναι λογικό να γίνει η εκπαίδευση του "προσεγγιστή" συνάρτησης με παραδείγματα από την κατανομή αυτή.
Μια κατανομή ιδιαίτερης σημασίας είναι **η κατανομή της πολιτικής** ,όπου η κατανομή αυτή περιέχει όλες τις καταστάσεις από τις οποίες περνά ο παίκτης ακολου-

θώντας την πολιτική αυτή. Με αυτόν τον τρόπο γίνεται προσπάθεια ελαχιστοποίησης σφάλματος των καταστάσεων αυτών αγνοώντας τις υπόλοιπες. Σύμφωνα με μελέτες, η χρήση της κατανομής αυτής αποφέρει καλύτερα αποτελέσματα σύγκλισης. Δεν είναι όμως απόλυτα σαφές το εάν πρέπει να μας απασχολεί μόνο η ελαχιστοποίηση του σφάλματος ελαχίστων τετραγώνων αφού ο τελικός στόχος είναι να χρησιμοποιήσουμε τις προβλέψεις για την εύρεση μιας καλύτερης πολιτικής. Οι βέλτιστες προβλέψεις για τον σκοπό αυτό δεν είναι απαραίτητα και οι καλύτερες για την ελαχιστοποίηση του σφάλματος ελαχίστων τετραγώνων. Παρ'όλα αυτά, δεν είναι ξεκάθαρο ποιος θα μπορούσε να είναι ένας καλύτερος στόχος για τις προβλέψεις αυτές.

Ένας ιδανικός στόχος για το σφάλμα ελαχίστων τετραγώνων θα ήταν η εύρεση ενός ολικού βέλτιστου, ένα διάνυσμα παραμέτρων $\overrightarrow{\theta^*}$ για το οποίο

$$MSE(\overrightarrow{\theta^*}) \leq MSE(\overrightarrow{\theta})$$

για κάθε πιθανό $\overrightarrow{\theta}$. Αυτός είναι ένα εφικτός στόχος για απλούς 'προσεγγιστές'(ψαξε το) συναρτήσεων όπως γραμμικούς. Ο στόχος αυτός είναι ανέφικτος για πολύπλοκους "προσεγγιστές" συναρτήσεων οι οποίοι έχουν ως στόχο την σύγκλιση σε ένα τοπικό βέλτιστο, δηλαδή ένα διάνυσμα παραμέτρων $\overrightarrow{\theta^*}$ για το οποίο

$$MSE(\overrightarrow{\theta^*}) \leq MSE(\overrightarrow{\theta})$$

για όλα τα $\overrightarrow{\theta}$ σε μια γειτονιά του $\overrightarrow{\theta^*}$. Μέχρι τώρα δεν έχει βρεθεί καλύτερος στόχος για τους πολύπλοκους αυτούς "προσεγγιστές".

3.2 Μέθοδοι προσέγγισης συναρτήσεων

Παρακάτω αναλύονται οι μέθοδοι προσέγγισης συναρτήσεων στις κατηγορίες τους και επεξηγούνται η κάθε μία ξεχωριστά:

Μέθοδοι κατάβασης κλίσης(gradient descent)

Στις μεθόδους αυτές, το διάνυσμα παραμέτρων είναι ένα διάνυσμα στήλης με ένα δεδομένο αριθμό στοιχείων με πραγματικές τιμές , $\overrightarrow{\theta_t} = (\theta_t(1), \theta_t(2), ..., \theta_t(n))^T$ και $V_t(s)$ είναι μια λεία διαφορική συνάρτηση του $\overrightarrow{\theta_t}$ για όλα τα $s \in S$. Προς το παρόν ας υποθέσουμε ότι σε κάθε χρονικό βήμα t παρατηρείται ένα νέο παράδειγμα όπου $s_t \mapsto V^\pi(s_t)$. Αν και οι ακριβείς και σωστές αξίες $V^\pi(s_t)$ δίδονται,το πρόβλημά μας παραμένει δύσκολο διότι ο προσεγγιστής συνάρτησης έχει περιορισμένες πηγές δεδομένων και συνεπώς έχει περιορισμένη ανάλυση στην επίλυσή του. Δεν υπάρχει γενικώς κάποιο $\overrightarrow{\theta}$ που να βρίσκει ακριβώς σωστά όλες τις καταστάσεις ή ακόμα και όλα τα παραδείγματα. Επιπλέον, πρέπει να γενικεύσουμε και για όλες τις άλλες καταστάσεις που δεν παρουσιάστηκαν στα παραδείγματα. Υποθέτουμε ότι οι καταστάσεις εμφανίζονται στα παραδείγματα με την ίδια κατανομή P πάνω στην οποία προσπαθούμε να ελαχιστοποιήσουμε το σφάλμα ελαχίστων τετραγώνων. Οι

μέθοδοι κατάβασης κλίσης προσπαθούν να ελαχιστοποιήσουν το σφάλμα στα παρατηρούμενα παραδείγματα με την αλλαγή του διανύσματος παραμέτρων προς την κατεύθυνση που θα ελαχιστοποιήσει περισσότερο το σφάλμα :

$$\overrightarrow{\theta_{t+1}} = \overrightarrow{\theta_t} - (1/2)\alpha \bigtriangledown_{\overrightarrow{\theta_t}} [V^{\pi}(s_t) - V_t(s_t)]^2 = \overrightarrow{\theta_t} + \alpha[V^{\pi}(s_t) - V_t(s_t)] \bigtriangledown_{\overrightarrow{\theta_t}} V_t(s_t)$$

όπου α είναι μια θετική βηματική παράμετρος και $\nabla_{\overrightarrow{\theta_t}} f(\overrightarrow{\theta_t})$ είναι για κάθε συνάρτηση f το διάνυσμα

$$(\vartheta f(\overrightarrow{\theta_t})/\vartheta\theta_t(1), \vartheta f(\overrightarrow{\theta_t})/\vartheta\theta_t(2), ..., \vartheta f(\overrightarrow{\theta_t})/\vartheta\theta_t(n))^T$$

. Το παραπάνω διάνυσμα είναι η κλίση της συνάρτησης f ως προς το $\overrightarrow{\theta_t}$. Για αυτό και οι μέθοδοι αυτοί ονομάζονται μέθοδοι κατάβασης κλίσης.
Γενικώς, όταν για διάφορους λόγους υπάρχει θόρυβος στον υπολογισμό της συνάρτησης αξιολογήσεων ,τότε μπορούμε να υπολογίσουμε μόνο μια προσέγγισή της. Επομένως, η μέθοδος καθόδου κατά βαθμίδα γενικεύεται με την αντικατάσταση της προσεγγιστικής συνάρτησης με την πραγματική και η αναδρομική εξίσωση για το διάνυσμα παραμέτρων θα είναι :

$$\overrightarrow{\theta_{t+1}} = \overrightarrow{\theta_t} + \alpha[u_t - V_t(s_t)]\nabla_{\overrightarrow{\theta_t}} V_t(s_t)$$

Εάν η u_t είναι μια εκτίμηση που δεν εξαρτάται από τις αρχικές τιμές, δηλαδή ισχύει $E\{u_t\} = V^{\pi}(s_t)$ για κάθε t τότε εγγυημένα το $\overrightarrow{\theta_t}$ θα συγκλίνει σε ένα τοπικό βέλτιστο υπό τις γνωστές συνθήκες στοχαστικής προσέγγισης για τη μείωση της παραμέτρου α.

Γραμμικές μέθοδοι

Μια από τις πιο σημαντικές ειδικές περιπτώσεις της προσέγγισης συνάρτησης βαθμίδας καθόδου είναι όταν η συνάρτηση προσέγγισης ,V_t, είναι μια γραμμική συνάρτηση του διανύσματος παραμέτρων, $\overrightarrow{\theta_t}$. Υπάρχει ένα διάνυσμα στήλης χαρακτηριστικών συναρτήσεων που ανταποκρίνεται στην κάθε κατάσταση, $\overrightarrow{\phi_s} = (\phi_s(1), \phi_s(2), ..., \phi_s(n))^T$ με τον ίδιο αριθμό στοιχείων με του $\overrightarrow{\theta_t}$.Οι χαρακτηριστικές συναρτήσεις αυτές κατασκευάζονται από τις καταστάσεις με διάφορους τρόπους. Με όποιο τρόπο και να κατασκευάζονται, η προσεγγιστική συνάρτηση κατάστασης-αξιολόγησης δίδεται από την εξής εξίσωση:

$$V_t(s) = \overrightarrow{\theta_t}^T \overrightarrow{\phi_s} = \sum_{i=1}^{n} \theta_t(i)\phi_s(i)$$

.Όμως, $\nabla_{\overrightarrow{\theta_t}} V_t(s) = \overrightarrow{\phi_s}$. Σε αυτή την περίπτωση υπάρχει μόνο ένα βέλτιστο $\overrightarrow{\theta^*}$. Η επιλογή των χαρακτηριστικών συναρτήσεων κατάλληλων για την επίτευξη του τελικού στόχου είναι ένας σημαντικός τρόπος πρόσθεσης εκ των προτέρων βασικών πληροφοριών σε συστήματα Ενισχυτικής Μάθησης. Τέλος, οι συναρτήσεις αυτές θα πρέπει να ανταποκρίνονται στα φυσικά χαρακτηριστικά του στόχου.

Τρόποι χρήσης χαρακτηριστικών συναρτήσεων

COARSE Κωδικοποίηση
Εφαρμόζεται για στόχους όπου ο χώρος των καταστάσεων είναι συνεχής και δισδιάστατος. Ένα είδος χαρακτηριστικής συνάρτησης για αυτήν την περίπτωση είναι ο κύκλος. Δηλαδή, εάν η κατάσταση βρίσκεται μέσα σε ένα κύκλο,τότε η αντίστοιχη χαρακτηριστική συνάρτηση έχει αξιολόγηση 1 και λέμε ότι είναι παρούσα αλλιώς έχει αξιολόγηση 0 και λέμε ότι είναι απούσα. Αυτού του είδους οι χαρακτηριστικές συναρτήσεις ονομάζονται δυαδικές. Οπότε, έχοντας δεδομένη μια κατάσταση, οι αξίες των χαρακτηριστικών συναρτήσεων υποδεικνύουν και την θέση της κατάστασης αυτής. Ανάλογα με το μέγεθος και το σχήμα των χαρακτηριστικών συναρτήσεων επιτυγχάνεται και η αντίστοιχη γενίκευση στις άλλες καταστάσεις με την ανάλογη ακρίβεια.
Κωδικοποίηση Πλακιδίων
Η κωδικοποίηση αυτή είναι μια μορφή coarse κωδικοποίησης που χρησιμεύει όταν απαιτείται αποτελεσματική μάθηση κατά την διάρκεια της αλληλεπίδρασης του παίκτη με το περιβάλλον του. Στην κωδικοποίηση αυτή τα πεδία ορισμού των χαρακτηριστικών συναρτήσεων ομαδοποιούνται σε αναλυτικά τμήματα του χώρου εισόδου. Κάθε τμήμα ονομάζεται tiling και κάθε στοιχείο του τμήματος αυτού ονομάζεται πλακίδιο. Κάθε πλακίδιο είναι το πεδίο ορισμού μιας δυαδικής χαρακτηριστικής συνάρτησης. Μόνο μια χαρακτηριστική συνάρτηση είναι παρούσα σε κάθε tiling, οπότε ο συνολικός αριθμός των χρησιμοποιούμενων χαρακτηριστικών συναρτήσεων είναι πάντα ο ίδιος με αυτόν των tilings. Ένα κόλπο για την μείωση των απαιτήσεων μνήμης είναι ο χωρισμός του tiling σε μικρότερα χωρίς να χάνεται πολύ πληροφορία και έτσι να απαιτείται υψηλή ανάλυση σε ένα μικρό τμήμα του χώρου καταστάσεων.

Μέθοδοι ελέγχου με τη χρήση γενικευμένης επανάληψης πολιτικής

Με τη χρήση του πρότυπου της γενικευμένης επαναληπτικής πολιτικής η μέθοδος λειτουργεί ως εξής: Πρώτον, οι μέθοδοι πρόβλεψης αξιολογήσεων καταστάσεων αντικαθίστανται από τις μεθόδους πρόβλεψης αξιολογήσεων ενεργειών και μετά συνδυάζονται με την βελτίωση πολιτικής και τις τεχνικές επιλογής ενεργειών. Η γενική εξίσωση ανανέωσης της κατάβασης κλίσης είναι :

$$\overrightarrow{\theta_{t+1}} = \overrightarrow{\theta_t} + \alpha[u_t - Q_t(s_t, \alpha_t)] \bigtriangledown_{\overrightarrow{\theta_t}} Q_t(s_t, \alpha_t)$$

3.3 Γενίκευση και πρόβλεψη αξιών μέσω των προσεγγιστικών συναρτήσεων στην περίπτωση του Tetris

Η ιδέα της γενίκευσης και πρόβλεψης αξιών μέσω των προσεγγιστικών συναρτήσεων, που αναλύθηκε ανωτέρω, εφαρμόστηκε στο παιχνίδι Tetris και αποτελεί τον πυρήνα της διπλωματικής εργασίας. Πιο συγκεκριμένα, στο παιχνίδι αυτό μελετήθη-

κε και αναλύθηκε το περιβάλλον του,δηλαδή καταγράφηκαν όλα τα χαρακτηριστικά που περιγράφουν και αποτελούν το περιβάλλον αυτό. Η μελέτη αυτή ήταν ζωτικής σημασίας διότι με αυτόν τον τρόπο λήφθησαν τα χαρακτηριστικά τα οποία αποτελούν το προαναφερθέν διάνυσμα παραμέτρων θ_t. Τα ανωτέρω χαρακτηριστικά ονομάζονται και χαρακτηριστικές συναρτήσεις (basis functions) και είναι τα εξής :

get landing height : η συνάρτηση αυτή επιστρέφει το ύψος του τελευταίου τούβλου που έπεσε,δηλαδή το ύψος στο οποίο βρίσκεται το υψηλότερο τμήμα του. Το ύψος αυτό ορίζεται από το μεσαίο τμήμα του τούβλου αυτού.

get eroded piece cells : η συνάρτηση αυτή αξιολογεί πόσο συνέβαλε το τελευταίο τούβλο στην αφαίρεση γραμμών.Επομένως, επιστρέφει έναν αριθμό τούβλων ο οποίος είναι $n*m$, όπου n είναι ο αριθμός των γραμμών που διαγράφηκαν και m είναι ο συνολικός αριθμός των τούβλων που διαγράφηκαν με την τελική τοποθέτηση του τούβλου αυτού.

get row transitions : η συνάρτηση αυτή επιστρέφει τις ανωμαλίες των γραμμών, δηλαδή στην ουσία τα κενά που υπάρχουν στην γραμμή αυτή. Η ονομασία της συνάρτησης προκύπτει από το ότι γνωρίζοντας τις ανωμαλίες αυτές γνωρίζουμε και τον αριθμό των μεταβάσεων της γραμμής αυτής προς τα δεξιά και τα αριστερά. Τα τοιχώματα του περιβάλλοντος του Tetris θεωρούνται ως γεμάτες γραμμές, οπότε μια άδεια γραμμή έχει δύο δυνατές μεταβάσεις.

get column transitions : η συνάρτηση αυτή επιστρέφει τις ανωμαλίες των στηλών,επομένως επιστρέφει τον αριθμό των δυνατών μεταβάσεων των στηλών με την ίδια λογική με την προηγούμενη συνάρτηση. Η κάτω οριακή γραμμή του περιβάλλοντος του Tetris θεωρούνται ως γεμάτες γραμμές, άρα μια άδεια στήλη έχει μόνο μια μετάβαση.

get holes : η συνάρτηση αυτή επιστρέφει τον αριθμό των οπών πάνω στον πίνακα του περιβάλλοντος. Η οπή είναι ένα άδειο κελί πάνω από το οποίο υπάρχει τουλάχιστον ένα γεμάτο κελί ,πάντοτε στην ίδια στήλη.

get well sums dellacherie : η συνάρτηση αυτή αξιολογεί τα πηγάδια και υπολογίζει πόσο βαθειά είναι. Τα πηγάδια ορίζονται ως οι θέσεις όπου μπορούν να πληρωθούν με ίσια τούβλα καθέτως. Ονομάζουμε ”κελί πηγαδιού” το άδειο κελί του οποίου το αριστερό και δεξί κελί είναι γεμάτο. Όσον αφορά την αξιολόγηση των πηγαδιών αυτών, ακολουθείται η εξής διαδικασία: σε κάθε στήλη, αναζητούνται τα ”κελιά πηγαδιού” και για κάθε κάθε τέτοιο κελί μετρώνται τα άδεια κελιά που βρίσκονται κάτω από το ”κελί πηγαδιού” συμπεριλαμβανομένου και του κελιού αυτού. Η αξιολόγηση κάθε πηγαδιού για να γίνει η αντίστοιχη αξιολόγηση είναι το άθροισμα όλων των βαθών όλων των ”κελιών πηγαδιών” που βρίσκονται στην ίδια στήλη. Με άλλα λόγια, το πηγάδι αποτελείται από όλα τα ”κελιά πηγαδιών” που βρίσκονται σε μια στήλη.

get wall height : η συνάρτηση αυτή επιστρέφει το ύψος του τείχους, το οποίο ορίζεται ως το ύψος που βρίσκεται η χαμηλότερη άδεια γραμμή.

get next column height : η συνάρτηση αυτή επιστρέφει το ύψος της κάθε στήλης. Άρα, η συνάρτηση αυτή πρέπει να κληθεί τόσες φορές όσος είναι το μέγεθος του

πλάτους του περιβάλλοντος. Κάθε φορά που καλείται, η συνάρτηση αυτή πάει στην επόμενη στήλη.

get next column height difference : η συνάρτηση αυτή επιστρέφει την διαφορά ύψους μεταξύ δύο διαδοχικών στηλών. Άρα, η συνάρτηση αυτή πρέπει να κληθεί τόσες φορές όσος είναι το μέγεθος του πλάτους του περιβάλλοντος μειωμένο κατά ένα εφόσον πρόκειται για σύγκριση των στηλών ανά δύο. Κάθε φορά που καλείται, η συνάρτηση αυτή πάει στο επόμενο ζεύγος στηλών.

get next wall height : η συνάρτηση αυτή επιστρέφει 1 εάν το ύψος του τείχους είναι n και 0 αλλιώτικα ,όπου n είναι ο αριθμός που έχουμε καλέσει την συνάρτηση αυτή στην συγκεκριμένη κατάσταση που βρισκόμαστε. Με άλλα λόγια, την πρώτη φορά που καλείται η συνάρτηση αυτή, επιστρέφει αυτή 1 εάν το ύψος του τείχους είναι 0 και 0 αλλιώτικα,την δεύτερη φορά θα επιστρέψει 1 εάν το ύψος είναι 1 και 0 αλλιώτικα και την n-ιοστή φορά θα επιστρέψει 1 εάν το ύψος είναι n-1 και 0 αλλιώτικα. Επομένως, η συνάρτηση πρέπει να καλεστεί τόσες φορές όσο είναι το μέγεθος του ύψους του πίνακα. Με τη συνάρτηση αυτή δύναται να κατηγοριοποιηθούν οι καταστάσεις του παιχνιδιού μας ως προς το ύψος του τείχους τους ώστε να μπορέσουμε να αντιστοιχίσουμε βάρη με κάθε ένα ύψος τείχους.

get hole depths : η συνάρτηση αυτή επιστρέφει το σύνολο των βαθών των οπών. Το βάθος της κάθε οπής είναι το σύνολο των γεματων κελιών πάνω από το άδειο κελί μας που αποτελεί και την οπή μας.

get rows with holes : η συνάρτηση αυτή επιστρέφει τον αριθμό των γραμμών με μια οπή.

get next local value function : η συνάρτηση αυτή επιστρέφει το ακριβές αποτέλεσμα της συνάρτησης αξιολόγησης ενός πίνακα 5*5 του περιβάλλοντος Tetris. Η συνάρτηση αυτή πρέπει να καλεστεί τόσες φορές όσοι είναι οι πιθανές θέσεις ενός 5*5 πίνακα στον δεδομένο μας πίνακα, όπου στην συγκεκριμένη περίπτωση για τον πίνακα μας που είναι 10*20 είναι 96 θέσεις. Κάθε φορά που κλείται η συνάρτηση αυτή πάει στην επόμενη θέση.

get well sums fast : η συνάρτηση αυτή επιστρέφει την αξιολόγηση και το βάθος των ”κελιών-πηγαδιών”. Η μόνη διαφορά με την αντίστοιχη συνάρτηση που προαναφέρθηκε είναι ότι εδώ όλα τα άδεια κελιά που βρίσκονται κάτω από το ”κελί-πηγάδι” θεωρούνται και αυτά ”κελιά-πηγάδια”.

get occupied cells : η συνάρτηση αυτή επιστρέφει τον αριθμό των γεμάτων κελιών του πίνακα.

get weighted cells : η συνάρτηση αυτή επιστρέφει τον αριθμό των γεμάτων κελιών του πίνακα σταθμισμένα με τα ύψη τους ως βάρη.

get wells : η συνάρτηση αυτή επιστρέφει τον αριθμό των ”κελιών-πηγαδιών”. Εδώ τα κελιά αυτά ορίζονται όπως λέει η συνάρτηση get well sums fast.

get row eliminated : η συνάρτηση αυτή επιστρέφει τον αριθμό γραμμών που διαγραφήθηκαν στην τελευταία κίνηση του κομματιού.

get max height difference : η συνάρτηση αυτή επιστρέφει την μέγιστη διαφορά ύψους μεταξύ των στηλών.

get wall distance to top : η συνάρτηση αυτή επιστρέφει την απόσταση του τείχους από το πάνω μέρος του πίνακα περιβάλλοντος. Η απόσταση αυτή ορίζεται από το πάνω αυτό μέρος και από το ύψος της χαμηλότερης άδειας γραμμής.
get next column distance to top : η συνάρτηση αυτή επιστρέφει την απόσταση των στηλών από το πάνω μέρος του πίνακα.
get next column height difference2 : η συνάρτηση αυτή επιστρέφει την διαφορά ύψους μεταξύ δύο διαδοχικών στηλών.
get wall distance to top square : η συνάρτηση αυτή επιστρέφει την απόσταση του τείχους από το πάνω μέρος του πίνακα υψωμένη στο τετράγωνο.
get height square : η συνάρτηση αυτή επιστρέφει το ύψος του τείχους(ελεγξε) υψωμένο στο τετράγωνο.
get hole depths square : η συνάρτηση αυτή επιστρέφει τα βάθη των οπών υψωμένα στο τετράγωνο.
get next column height2 : η συνάρτηση αυτή επιστρέφει τα ύψη των στηλών.
get diversity : η συνάρτηση αυτή επιστρέφει την διαφορετικότητα της κορυφής του τείχους,με άλλα λόγια την ανομοιομορφία της άνω γραμμής του τείχους αυτού.

Συνάρτηση αξιολογήσεων και στρατηγικές παιξίματος

Αυτές λοιπόν αποτελούν το σύνολο των βασικών συναρτήσεων που περιγράφουν όσο καλύτερα γίνεται το παιγνίδι του Tetris. Από αυτές επιλέχθηκαν συγκεκριμένες για να δημιουργηθεί η συνάρτηση αξιών μέσω της γραμμικής μεθόδου

$$V_t(s) = \overrightarrow{\theta_t}^T \overrightarrow{\phi_s} = \sum_{i=1}^{n} \theta_t(i)\phi_s(i)$$

και αυτές είναι : landing height, eroded piece cells, row transitions, column transitions, hole number, well sums, hole depths, rows with holes, max height, diversity. Ο τρόπος αυτός της δημιουργίας της συνάρτησης αξιών ονομάζεται γραμμική προσέγγιση της πραγματικής άγνωστης συνάρτησης αξιών. Στο συγκεκριμένο πρόβλημα που έπρεπε να επιλυθεί χρησιμοποιήθηκαν δύο στρατηγικές, η μία που χρησιμοποιεί όλες τις παραπάνω βασικές συναρτήσεις και η άλλη που χρησιμοποιεί τυχαία διανύσματα.

Η μέθοδος cross entropy και η εφαρμογή της στο παιχνίδι Tetris

Η μέθοδος **cross entropy** είναι ένας γενικός αλγόριθμος για την προσεγγιστική επίλυση προβλημάτων ολικής βελτιστοποίησης της μορφής

$$w^* = arg\max_w S(w)$$

όπου S είναι μια γενικευμένη αντικειμενική συνάρτηση (δηλαδή δεν χρειάζεται να υποτεθεί η συνέχειά της). Ενώ οι περισσότεροι αλγόριθμοι βελτιστοποίησης δίδουν μιας απλής μορφής υποψήφια λύση σε κάθε χρονικό βήμα, όπως w_t, η μέθοδος που

χρησιμοποιείται δίδει σε κάθε χρονικό βήμα μια κατανομή πιθανών λύσεων και την ανανεώνει ομοιοτρόπως.

Η διαδικασία της μεθόδου αυτής είναι η εξής : Αρχικά, το διάνυσμα των παραμέτρων ή βαρών είναι ένα διάνυσμα με μηδενικά και μας δίδεται η κατανομή των δειγμάτων από την οποία θα ληφθούν π.χ. Gaussian με γνωστή την μέση τιμή και την διασπορά της. Από την κατανομή αυτή επιλέγονται τόσα δείγματα όσα είναι και τα στοιχεία του διανύσματος παραμέτρων και με το νεοσύστατο διάνυσμα εκτελείται το πείραμα, το οποίο είναι στην ουσία η εκτέλεση του κώδικα με τις προαναφερόμενες τιμές των παραμέτρων που του δόθηκαν. Μετά την εκτέλεση αυτή επιστρέφεται η συνολική ανταμοιβή για το πείραμα αυτό και η διαδικασία αυτή επαναλαμβάνεται όσες φορές έχουν δοθεί ως όρισμα στον κώδικα, έστω m φορές. Έπειτα, αυτές οι m ανταμοιβές συγκρίνονται μεταξύ τους και επιλέγονται οι $\rho * m$ καλύτερες, όπου ρ είναι το ποσοστό που έχει δοθεί από τον χρήστη και αυτό το ποσοστό υποδηλώνει τις πόσες καλύτερες ανταμοιβές θα επιλεχθούν. Για το ρ ισχύει 0<ρ<1. Η επιλογή αυτή γίνεται διότι από αυτές τις επιλεχθήσες ανταμοιβές καθορίζονται και ανανεώνονται παράλληλα οι μέσες τιμές και οι διασπορές των νέων κατανομών από τις οποίες επιλέγονται τα νέα δείγματα και η διαδικασία επαναλαμβάνεται από την αρχή. Οι τύποι μέσω των οποίων προκύπτουν οι νέες κατανομές είναι οι εξής :

$$\mu_{t+1} := \frac{\sum_{i \in I} w_i}{\mid I \mid}$$

$$\sigma^2_{t+1} := \frac{\sum_{i \in I} (w_i - \mu_{t+1})^T (w_i - \mu_{t+1})}{\mid I \mid}$$

Στους παραπάνω τύπους το $\mid I \mid$ είναι ο αριθμός των βαρών. Ένας από τους λόγους που χρησιμοποιείται η μέθοδος αυτή είναι ότι σε κάθε ανανέωση της μέσης τιμής και της διασποράς από τους παραπάνω τύπους, στην νέα καμπύλη που δημιουργείται παρατηρείται ότι είναι πιο ”κλειστή” σε σχέση με την προηγούμενή της, δηλαδή το άνοιγμα της καμπύλης μικραίνει με την πάροδο του χρόνου. Έτσι, αποκτάται και η επιθυμητή σύγκλιση στο σημείο που ζητείται. Επομένως, σύμφωνα με τα παραπάνω που ειπώθηκαν, η μέθοδος cross entropy χρησιμοποιείται για την μάθηση των βαρών τα οποία λαμβάνονται από κατανομές Gauss. Ένα θέμα το οποίο απαιτεί αρκετή προσοχή και έχει γίνει θέμα έρευνας είναι ότι η εφαρμογή της μεθόδου αυτής σε προβλήματα Ενισχυτικής Μάθησης είναι αρκετά περιορισμένη διότι παρατηρείται ότι συγκλίνει σε ένα σημείο πολύ γρήγορα. Αυτό βεβαίως είναι κάτι το μη επιθυμητό επειδή ο παίκτης δεν προλαβαίνει να μάθει εγκαίρως. Για την αντιμετώπιση αυτού του προβλήματος υλοποιείται το εξής τέχνασμα που χρησιμοποιείται σε φιλτράρισμα σωματιδίων : σε κάθε επανάληψη ή ανανέωση της τιμής της διασποράς, προστίθεται κάποιος έξτρα θόρυβος στην κατανομή. Επομένως, ο τύπος υπολογισμού της γίνεται :

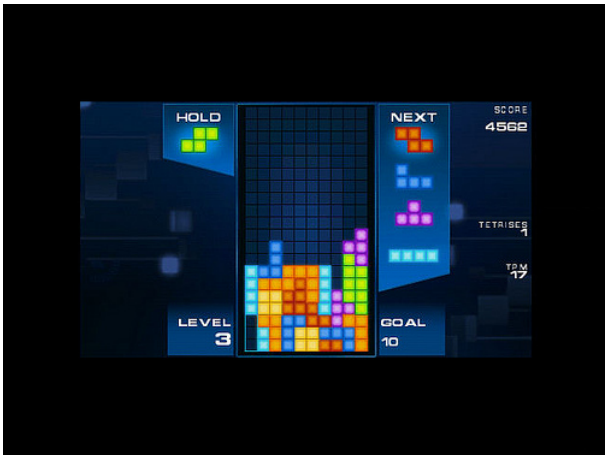
$$\sigma^2_{t+1} := \frac{\sum_{i \in I} (w_i - \mu_{t+1})^T (w_i - \mu_{t+1})}{\mid I \mid} + Z_{t+1}$$

, όπου Z ο θόρυβος. Άρα θόρυβος στην ουσία είναι μια ”αλλαγή” στην πραγματική τιμή της διασποράς που προκύπτει σε κάθε ανανέωση. Ο λόγος που ο θόρυβος αυτός προστίθεται είναι ότι είναι επιθυμητή η αύξηση της τιμής της διασποράς αφού αυτό σημαίνει ότι η καμπύλη ”ανοίγει” και έτσι θα συγκλίνει πιο αργά.

Κεφάλαιο 4

Πειράματα

Το πρόβλημα το οποίο πραγματεύεται της διπλωματικής εργασίας είναι η εφαρμογή του αλγορίθμου cross entropy πάνω στο ηλεκτρονικό παιχνίδι Tetris και η εξακρίβωση για το εάν ο αλγόριθμος αυτός είναι ο κατάλληλος για την σωστή και γρήγορη μάθηση του παίκτη. Όπως προαναφέρθηκε, το Tetris έχει γίνει θέμα μελέτης στο πεδίο της Ενισχυτικής Μάθησης για τους εξής λόγους : αποτελεί ένα πρόβλημα υψηλής πολυπλοκότητας το οποίο δεν έχει έναν προκαθορισμένο από τους κανόνες του παιχνιδιού τερματικό ορίζοντα και επίσης η εύρεση της βέλτιστης στρατηγικής είναι δύσκολη ακόμα και αν είναι γνωστή από πριν η ακολουθία των κομματιών που θα πέσουν. Επιπρόσθετα, οι αλγόριθμοι Ενισχυτικής Μάθησης είναι αρκετά αποτελεσματικοί στην επίλυση μιας ποικιλίας προβλημάτων αλυσιδωτών αποφάσεων, κατηγορία στην οποία εμπίπτει και το παιχνίδι Tetris, αιτιολογώντας και την επιλογή του αλγορίθμου μάθησης. Επίσης, ενώ οι περισσότεροι αλγόριθμοι βελτιστοποίησης κρατούν μία υποψήφια βέλτιστη λύση σε κάθε επανάληψη,η cross entropy κρατά μια κατανομή πιθανών λύσεων την οποία ανανεώνει σε κάθε επανάληψη.
Το πειραματικό μέρος για την περάτωση της παραπάνω εφαρμογής σχεδιάστηκε ως εξής : Για την εκτέλεση των πειραμάτων αυτών χρησιμοποιήθηκε ένας αρκετά εκτενής κώδικας σε γλώσσα προγραμματισμού Python. Ο κώδικας αυτός, ο οποίος αποτελεί και το βασικό κομμάτι του προγραμματισμού του παιχνιδιού Tetris, συνδέεται με άλλα αρχεία σε γλώσσα προγραμματισμού C και ο τρόπος σύνδεσής τους συνίσταται στο ότι από το βασικό αρχείο του κώδικα καλούνται τα υπόλοιπα αυτά αρχεία μέσω συγκεκριμένων συναρτήσεων. Στην ουσία, το βασικό αρχείο είναι το παιχνίδι Tetris, το οποίο αναπαράγει το διαδραστικό περιβάλλον, την πορεία του παιχνιδιού και τα άλλα χαρακτηριστικά του. Για την εξαγωγή των αποτελεσμάτων του παιχνιδιού υπάρχουν τρεις στρατηγικές-τρόποι παιξίματος. Η πρώτη καλεί δύο συναρτήσεις μέσα στο βασικό αρχείο του κώδικα ώστε να ενεργοποιηθεί ο αλγόριθμος cross entropy, που με την σειρά του θα δώσει το διάνυσμα βαρών **w** και το οποίο διάνυσμα θα ενσωματωθεί στον μαθηματικό τύπο υπολογισμού της συνάρτησης αξιών. Ως ενσωμάτωση εννοείται ότι λαμβάνονται τα στοιχεία του διανύσματος βαρών και τοποθετούνται στις αντίστοιχες θέσεις των συντελεστών των βασικών συναρτήσεων στο μαθηματικό τύπο της συνάρτησης αξιολογήσεων. Η δεύτερη στρατηγική,

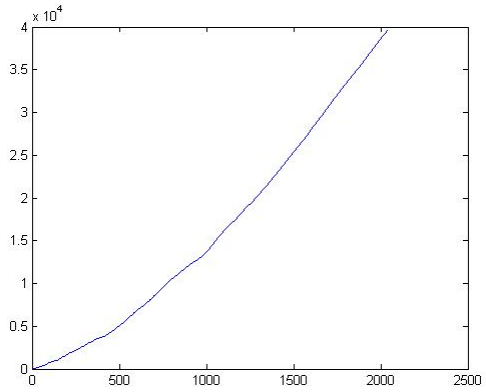


Σχήμα 4.1: Μια καλή στρατηγική.

στον μαθηματικό τύπο υπολογισμού της συνάρτησης αξιολογήσεων, έχει στη θέση των στοιχείων του διανύσματος βαρών προκαθορισμένες τυχαίες τιμές καθιστώντας την στρατηγική αυτή μια τυχαία πολιτική παιξίματος. Η τρίτη στρατηγική είναι όμοια με την προηγούμενη ως προς το ότι οι συντελεστές των βασικών συναρτήσεων είναι τυχαίοι αριθμοί,αλλά διαφέρει στο ότι σε κάθε επανάληψη οι αριθμοί αυτοί αλλάζουν τυχαία. Η λογική της χρήσης μιας τέτοιας μεθόδου είναι ότι εφόσον οι συντελεστές της αντικειμενικής συνάρτησης παράγονται τυχαία υπάρχει η προσδοκία εύρεσης του επιθυμητού βέλτιστου και της επίλυσης του προαναφερόμενου προβλήματος. Το πρόβλημα όμως έγκειται στο ότι όταν η επίλυση ενός προβλήματος εξαρτάται από την τυχαιότητα σε μεγάλο βαθμό, τότε αυτή η επίλυση θα πραγματοποιηθεί σε ένα χρονικό ορίζοντα για τον οποίο κανείς δεν μπορεί να ξέρει ποιος είναι.

Στη παρούσα εργασία η χρησιμοποιούμενη μέθοδος μάθησης αρχικοποιείται με την εισαγωγή τιμών της μέσης τιμής και διασποράς όπου στην συγκεκριμένη περίπτωση έχουν τεθεί ως μηδέν και 15 αντίστοιχα με την διασπορά ηθελημένα μεγάλη ώστε να δοθούν στο πρόγραμμα μεγάλα εύρη επιλογής υποψήφιων διανυσμάτων, πράγμα σημαντικό για προβλήματα μάθησης. Σε καθε επανάληψη λαμβάνονται 100 δείγματα-διανύσματα. Ο αριθμός αυτός επιλέχθηκε διότι παρατηρήθηκε ότι σε μικρότερους αριθμούς επαναλήψεων η μάθηση δεν εξελισσόταν επιθυμητά, δηλαδή καθυστερούσε πολύ η πρόοδος του παίκτη και συχνά λόγω αυτής της έλλειψης μάθησης εγκλωβιζόταν σε τοπικά βέλτιστα. Η αξιολόγηση του κάθε δείγματος ώστε να μπορεί να συγκριθεί με τα υπόλοιπα 99 της παρτίδας(επανάληψης),γίνεται με την εκτέλεση ενός παιχνιδιού Tetris χρησιμοποιώντας το συγκεκριμένο δείγμα. Άρα, το πρόγραμμα λαμβάνει αυτό το διάνυσμα και με την όποια μάθηση έχει αποκτήσει στο επίπεδο που βρίσκεται εκτελεί το παιχνίδι μέχρι να τερματιστεί και το τελικό σκορ που διαμορφώνεται αποτελεί και την τελική αξιολόγηση του δείγματος. Όπως σε όλα τα προβλήματα μάθησης, έτσι και εδώ υπάρχει ένα κριτήριο σύγκλισης, δηλαδή ένα κριτήριο τερματισμού των πειραμάτων και ολοκλήρωσης της εργασίας.

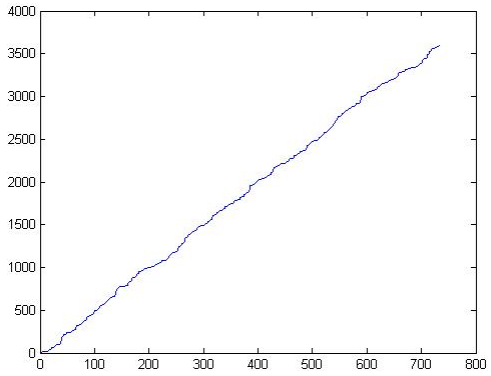
Το κριτήριο αυτό έχει ορισθεί ως ένα κατώτατο όριο της τιμής των διασπορών των



Σχήμα 4.2: Αποτελέσματα των συσσωρευμένων ανταμοιβών ως προς τον αριθμό των κινήσεων για την cross entropy.

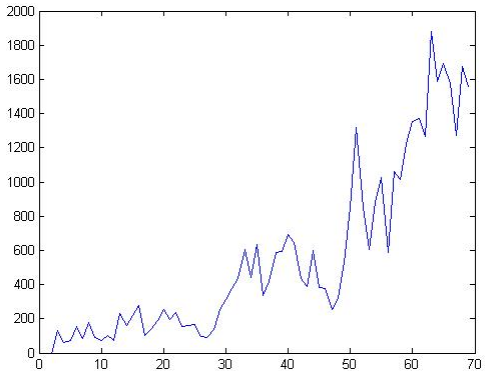
καμπυλών από όπου λαμβάνονται τα δείγματα της κάθε επανάληψης. Η επιλογή του κριτηρίου βασίστηκε στο ότι εάν οι τιμές των διασπορών γίνουν αρκετά μικρές σημαίνει ότι για τις παραμέτρους στις οποίες αντιστοιχούν αυτές οι καμπύλες δεν τίθεται περαιτέρω βελτίωση ως προς την συνεισφορά τους στην αξιολόγηση του δείγματος. Στην παρούσα εργασία το όριο αυτό έχει ορισθεί ώστε το άθροισμα όλων των διασπορών να μην ξεπερνά την τιμή 1. Όσο αφορά τις ποσότητες που απεικονίζονται διαγραμματικά ως προς τον συνολικό χρόνο εκτέλεσης των πειραμάτων,θα αναλυθεί παρακάτω ο τρόπος υπολογισμού τους. Καταρχάς, η συνολική ανταμοιβή που πήρε ο παίκτης μετά το πέρας των επαναλήψεων υπολογίστηκε με τον εξής τρόπο. Μετά από την ολοκλήρωση της κάθε κίνησης του παίκτη το πρόγραμμα δίδει στον χρήστη μια ανταμοιβή σε μορφή ενός αριθμού μετά την τελική τοποθέτηση του τούβλου μέσα στον χώρο του παιχνιδιού. Επομένως, ανάλογα με το πόσο καλή ήταν η κίνηση αυτή το πρόγραμμα θα του δώσει και ανάλογη ανταμοιβή. Μια κίνηση ορίζεται ως «καλή» συμφωνα με το Tetris και τους κανόνες του. Στον κώδικα καταχωρείται μετά από κάθε τέτοια κίνηση η τιμή της αντίστοιχης ανταμοιβής σε ένα διάνυσμα έχοντας τελικά μέγεθος ίδιου με τον αριθμό όλων των κινήσεων του παίκτη κατά την διάρκεια των επαναλήψεων που εκτελέστηκαν. Ακολούθως αθροίζονται τα στοιχεία του διανύσματος αυτού μεταξύ τους ανά 1000 και τοποθετούνται σε διάγραμμα ως προς τον αριθμό των κινήσεων που αντιστοιχούν στις ανταμοιβές αυτές όπως φαίνεται στα παραπάνω διαγράμματα:

Στο σχήμα 4.2 και 4.3 παρατηρείται ότι στην αρχή οι ανταμοιβές που δίδονται από το παιχνίδι είναι αρκετά χαμηλές, δηλαδή ο παίκτης δεν έχει ακόμα ορθή γνώση του παιχνιδιού όπως είναι αναμενόμενο. Με την πάροδο όμως των επαναλήψεων παρατηρείται μια μεγάλη αύξηση των συσσωρευμένων ανταμοιβών που αποδεικνύει ότι έχει λάβει χώρα η μάθηση με αυξανόμενο ρυθμό κυρίως από το χρονικό βήμα 490,000 περίπου και πέρα. Τούτο συμβαίνει διότι η επιλογή της μεθόδου μάθησης καθοδήγησε κατάλληλα τον παίκτη κατά την διάρκεια των επαναλήψεων ώστε να

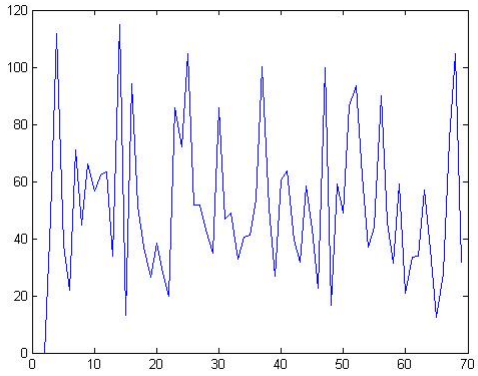


Σχήμα 4.3: Αποτελέσματα των συσσωρευμένων ανταμοιβών ως προς τον αριθμό των κινήσεων για την τυχαία πολιτική.

φτάσει στο επιθυμητό επίπεδο. Αξίζει να σημειωθεί ότι ο παίκτης κατάφερε να παίξει παιχνίδια για περισσότερα χρονικά βήματα όταν ακολουθούσε την μέθοδο μάθησης σε σχέση με όταν ακολουθούσε την τυχαία πολιτική με αποτέλεσμα να συλλέξει και περισσότερη ανταμοιβή. Αντιθέτως, παρατηρείται ότι, όπως είναι αναμενόμενο, στην περίπτωση της τυχαίας πολιτικής η κλίση της καμπύλης συσσωρευμένης ανταμοιβής ως προς τα χρονικά βήματα είναι περίπου σταθερή. Επίσης, για να εκτιμηθεί καλύτερα η απόδοση της μεθόδου μάθησης χρησιμοποιήθηκε σαν δείκτης η ανταμοιβή που συγκέντρωσε ο παίκτης σε κάθε παίξιμο και η μεταβολή της τιμής της με το χρόνο των παιξιμάτων διότι η ανταμοιβή ανά επεισόδιο δείχνει με την μεγαλύτερη δυνατή σαφήνεια την πορεία της μάθησης με το χρόνο. Από τα διαγράμματα των ανταμοιβών ανά επανάληψη για τις δύο πολιτικές παιξίματος εξάγονται τα εξής συμπεράσματα: στο διάγραμμα της τυχαίας πολιτικής, φαίνεται ότι η ανταμοιβή ανά επεισόδιο κυμαίνεται μεταξύ των ίδιων περίπου άνω και κάτω ορίων που είναι 110 και 20 αντίστοιχα κατά την εκτέλεση όλων των επεισοδίων. Στο διάγραμμα της μεθόδου cross entropy είναι εμφανές ότι από ένα σημείο και μετά ο παίκτης έχει πλέον μάθει και έχει βελτιώσει αισθητά το επίπεδο παιξίματός του λαμβάνοντας και τις ανάλογα υψηλές ανταμοιβές ανά επεισόδιο ενώ πριν από αυτό λόγω της απειρίας του λαμβάνει αρκετά μικρές ανταμοιβές. Το σημείο αυτό προσδιορίζεται περί το επεισόδιο 2900. Περαιτέρω αισθητές βελτιώσεις παρατηρούνται σε περισσότερα του ενός σημεία, είναι δηλαδή «βηματική» η μάθηση και τα σημεία αυτά είναι περί τα επεισόδια 4850 και 5800,με το πρώτο σημείο να παρουσιάζει την μεγαλύτερη αύξηση της τιμής της ανταμοιβής που είναι προσεγγιστικά της τάξης του 700 τοις εκατό έναντι 300 τοις εκατό των άλλων δύο ως προς το αρχικό επίπεδο. Οι παραπάνω παρατηρήσεις επιβεβαιώνονται και από τα διαγράμματα του τρίτου δείκτη μάθησης που χρησιμοποιήθηκε και ο οποίος είναι ο αριθμός των χρονικών βημτων ανά επεισόδιο με την πάροδο του χρόνου παιξίματος. Μια παρατήρηση που ισχύει και για τις δύο τελευταίους δείκτες είναι ότι στην τυχαία πολιτική λόγω της μαθηματικής της

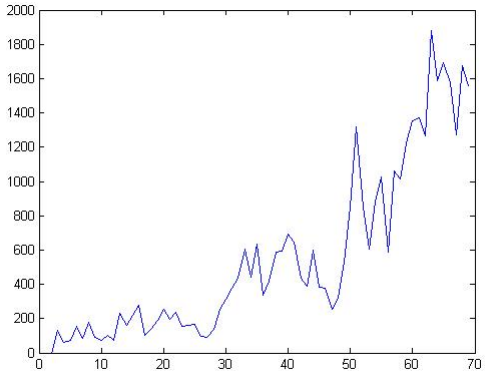


Σχήμα 4.4: Αποτελέσματα των συσσωρευμένων ανταμοιβών ανά επανάληψη για την cross entropy.

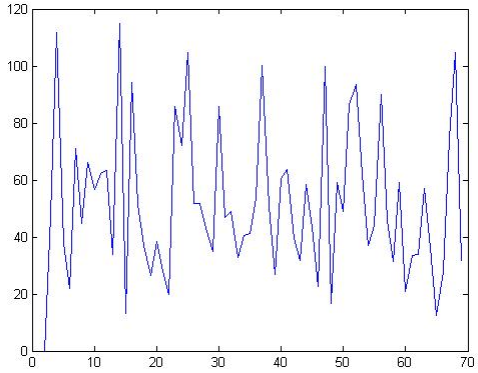


Σχήμα 4.5: Αποτελέσματα των συσσωρευμένων ανταμοιβών ανά επανάληψη για την τυχαία πολιτική.

φύσης υπάρχουν μεγάλες αποκλίσεις από την μέση τιμή των τιμών των μεταβλητών αυτών καθ'όλη την διάρκεια των επεισοδίων ενώ στην cross entropy οι αποκλίσεις αυτές υπάρχουν αλλά είναι κατά πολύ μικρότερες και με την πάροδο του χρόνου φαίνεται να αποσβένουν ελαφρώς.



Σχήμα 4.6: Αποτελέσματα των κινήσεων ανά επανάληψη για την cross entropy.



Σχήμα 4.7: Αποτελέσματα των κινήσεων ανά επανάληψη για την τυχαία πολιτική.

Κεφάλαιο 5

Συμπεράσματα και προτάσεις περαιτέρω μελέτης και βελτίωσης

Συμπεράσματα

Τα συμπεράσματα που εξήχθησαν από την παρούσα μελέτη είναι τα εξής : το ηλεκτρονικό παιχνίδι Tetris είναι ένα αρκετά πολύπλοκο παιχνίδι στρατηγικής με πολλά χαρακτηριστικά που το περιγράφουν και άρα είναι ως αντικείμενο μελέτης ιδιαίτερα απαιτητικό. Οι παράμετροι αυτές ονομάζονται χαρακτηριστικές συναρτήσεις και λαμβάνουν συγκεκριμένες τιμές η κάθε μία. Έπειτα, η Ενισχυτική Μάθηση με τη σειρά της είναι ένας τομέας ιδιαίτερα απαιτητικός και με πολλές μελλοντικές προεκτάσεις όσων αφορά τις μεθόδους της και σαν γενικό στόχο έχει την ταχύτερη και καλύτερη σύγκλιση της πολιτικής του παίκτη στην βέλτιστη λύση με την επιλογή των κατάλληλων ενεργειών στον κατάλληλο χρόνο. Όσον αφορά τον αλγόριθμο cross entropy που χρησιμοποιήθηκε στην διπλωματική εργασία, επρόκειτο για έναν αλγόριθμο επαναληπτικό που σκοπό έχει την εξαγωγή των καλύτερων βαρών των χαρακτηριστικών συναρτήσεων του παιχνιδιού Tetris ως προς την πολιτική, δηλαδή την μεγιστοποίηση της συνάρτησης αξιών και της ανταμοιβής. Επίσης, ενώ οι περισσότεροι αλγόριθμοι βελτιστοποίησης κρατούν μία υποψήφια βέλτιστη λύση σε κάθε επανάληψη, η cross entropy κρατά μια κατανομή πιθανών λύσεων την οποία ανανεώνει σε κάθε επανάληψη. Ως τελικό συμπέρασμα προκύπτει ότι ο αλγόριθμος cross entropy είναι κατάλληλος για την γρήγορη επίλυση προβλημάτων ίδιας πολυπλοκότητας με το πρόβλημα του Tetris.Άρα,με το συμπέρασμα αυτό επιβεβαιώνεται η καταλληλότητα του αλγορίθμου για το παιχνίδι αυτό [Istvan Szita et Andras Lorincz,2006]. Από τα αποτελέσματα της εφαρμογής του αλγορίθμου προκύπτει το συμπέρασμα ότι με την επιλογή του κατάλληλου αριθμού των δειγμάτων ανά επανάληψη η μάθηση έλαβε χώρα γρήγορα ο οποίος πρέπει γενικά να είναι μεγάλος έως ένα σημείο διότι πολύ μεγάλος αριθμός δειγμάτων συνεπάγεται καθυστέρηση στην εκτέλεση των επαναλήψεων και άρα και στην μάθηση.

Προτάσεις και βελτιώσεις

Αυτό το οποίο δίδει αρκετές ελπίδες βελτίωσης του αλγορίθμου cross entropy είναι ο πειραματισμός για την εύρεση των σωστών παραμέτρων καθορισμού του θορύβου της μεθόδου αυτής. Ως σωστοί εννοείται ως προς τον ορθό και εύχρηστο μαθηματικό τύπο και την ορθή επιλογή της τιμής τους. Βέβαια, υπάρχει και ένας κίνδυνος που είναι ο μεγάλος χρόνος εκτέλεσης του αλγορίθμου. Επίσης, προτείνεται από τους [Bohm et al.] ότι η χρήση εκθετικής συνάρτησης αξιών αντί του γραμμικού τύπου της δίδει μεγάλες βελτιώσεις. Τέλος και αυτό έχει πολύ μεγάλη σημασία, η προσεκτική και ορθή επιλογή και κατασκευή των χαρακτηριστικών συναρτήσεων συμβάλλει πολύ στην βελτίωση της απόδοσης οποιασδήποτε μεθόδου χρησιμοποιείται σε οποιοδήποτε πρόβλημα Ενισχυτικής Μάθησης.

Κεφάλαιο 6

Βιβλιογραφία

R. S. Sutton and A. G. Barto , Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998

Bohm, N., Kokai, G., and Mandl, S. , Evolving a heuristic function for the game Tetris, Proc. Lerner, Wissensentdeckung und Adaptivitat LWA, 2004

Istvan Szita et Andras Lorincz, Learning Tetris Using the Noisy Cross-Entropy Method, Department of Information Systems, Eotvos Lorand University , 2006

Demaine et al., 2003] Demaine, E. D., Hohenberger, S., and Liben-Nowell, D. (2003). Tetris is hard, even to approximate. In Proc. 9th International Computing and Combinatorics Conference (COCOON 2003), pages 351–363

[Farias nad van Roy,2006],Farias,V.F and van Roy,B.(2006).Probabilistic and Randomized Methods for Design Under Uncertainty,chapter Tetris:A Study of Randomized Constraint Sampling.Springer-Verlag UK.