

Πανεπιστήμιο Πατρών
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών

Τελική Έκθεση
για την εργασία του μαθήματος
Ευφυής Βιομηχανικός Έλεγχος

Reinforcement Learning

Θωμάς Ρεπαντής
Έτος: Ε΄
Κύκλος Σπουδών: Ηλεκτρονική και Υπολογιστές
Α.Μ.: 4218
E-mail: darkzero@otenet.gr

Ευαγγελία Κομισσοπούλου
Έτος: Ε΄
Κύκλος Σπουδών: Ηλεκτρονική και Υπολογιστές
Α.Μ.: 4157
E-mail: evacomihotmail.com

Πάτρα, 1.7.2002

1 Εισαγωγή στην επιβραβευμένη μάθηση

Μία από τις πρώτες απλές ιδέες σχετικά με τη μηχανική μάθηση ήταν και αυτή ενός συστήματος εκμάθησης που θέλοντας κάτι προσαρμόζει τη συμπεριφορά του προκειμένου να μεγιστοποιήσει ένα σήμα από το περιβάλλον του. Η ιδέα αυτή, ενός “ ηδονιστικού ” συστήματος εκμάθησης είναι και η βασική ιδέα της *επιβραβευμένης μάθησης* (ή *επιβραβεύομενης μάθησης* για να δηλώσουμε τη συνέχεια της διαδικασίας) (*Reinforcement Learning - RL*). Η επιβραβευμένη μάθηση βασίζεται στην ιδέα ότι μαθαίνει κανείς αλληλεπιδρώντας με το περιβάλλον του, καθώς προσπαθεί να επιτύχει κάποιο στόχο του. Πιο συγκεκριμένα αναφέρεται κυρίως σε μία υπολογιστική προσέγγιση της ιδέας αυτής.

2 Ιστορία της επιβραβευμένης μάθησης

Η ιστορία της επιβραβευμένης μάθησης αποτελείται από δύο κύριους κλάδους που αναπτύχθηκαν ξεχωριστά, πριν τελικά συνδυαστούν. Ο ένας κλάδος αφορά τη μάθηση μέσω δοκιμής και λάθους και ξεκίνησε από την ψυχολογία της μάθησης των ζώων και την τεχνητή νοημοσύνη. Ο άλλος κλάδος αφορά το πρόβλημα του βέλτιστου ελέγχου και τη λύση του χρησιμοποιώντας συναρτήσεις τιμών και δυναμικό προγραμματισμό. Οι δύο αυτοί αρχικά ανεξάρτητοι κλάδοι συνδυάστηκαν με έναν ακόμη που αφορά μεθόδους χρονικής διαφοράς. Και οι τρεις μαζί σχημάτισαν στα τέλη της δεκαετίας του 1980 την περιοχή της επιβραβευμένης μάθησης.

Μία από τις ισχυρότερες σύγχρονες τάσεις αφορά τη στενότερη επαφή μεταξύ της τεχνητής νοημοσύνης και άλλων μηχανικών επιστημών. Οι περιοχές μεταξύ της τεχνητής νοημοσύνης και της συμβατικής επιστήμης των μηχανικών είναι σήμερα ιδιαίτερα δραστήριες και περιλαμβάνουν πεδία όπως τα νευρωνικά δίκτυα, ο ευφυής έλεγχος και η επιβραβευμένη μάθηση. Στην επιβραβευμένη μάθηση επεκτείνονται ιδέες από το βέλτιστο έλεγχο και τη στοχαστική προσέγγιση για να αντιμετωπισθούν οι πιο ευρείς και φιλόδοξοι στόχοι της τεχνητής νοημοσύνης.

3 Ορισμός της επιβραβευμένης μάθησης

Η επιβραβευμένη μάθηση αναφέρεται στη μάθηση του τι πρέπει να κάνει κάποιος - πως δηλαδή πρέπει να αντιστοιχίσει καταστάσεις σε πράξεις - προκειμένου να

μεγιστοποιήσει ένα σήμα αριθμητικής ανταμοιβής. Κανείς δεν υποδεικνύει στο μαθητευόμενο πως να πράξει, αντίθετα από ό,τι συμβαίνει στις περισσότερες μορφές μηχανικής μάθησης. Αντιθέτως, ο μαθητευόμενος πρέπει να ανακαλύψει ποιες πράξεις αποφέρουν τη μεγαλύτερη ανταμοιβή δοκιμάζοντάς τις. Στις πιο ενδιαφέρουσες περιπτώσεις οι πράξεις πιθανά δεν επηρεάζουν μόνο την άμεση ανταμοιβή, αλλά και την επόμενη κατάσταση και έτσι όλες τις ακόλουθες ανταμοιβές. Τα δύο αυτά χαρακτηριστικά -η έρευνα μέσω δοκιμής και λάθους και η καθυστερημένη ανταμοιβή- είναι τα πλέον σημαντικά και ξεχωριστά στοιχεία της επιβραβευμένης μάθησης.

Η επιβραβευμένη μάθηση καθορίζεται χαρακτηρίζοντας προβλήματα και όχι αλγόριθμους μάθησης. Οποιοσδήποτε αλγόριθμος είναι κατάλληλος για την επίλυση ενός τέτοιου προβλήματος μπορεί να θεωρηθεί αλγόριθμος επιβραβευμένης μάθησης. Η βασική ιδέα είναι η σύλληψη των πιο σημαντικών όψεων του πραγματικού προβλήματος που αντιμετωπίζει ένας μαθητευόμενος πράκτορας που αλληλεπιδρά με το περιβάλλον του προκειμένου να επιτύχει ένα στόχο. Ένας τέτοιος πράκτορας πρέπει να είναι σε θέση να ανιχνεύει την κατάσταση του περιβάλλοντός του σε κάποιο βαθμό και επίσης να δρα ώστε να επηρεάζει την κατάσταση αυτή. Επιπλέον πρέπει να έχει έναν ή περισσότερους στόχους σχετιζόμενους με την κατάσταση του περιβάλλοντος. Οι τρεις αυτοί παράγοντες - η ανίχνευση του περιβάλλοντος, η δράση και ο στόχος - είναι οι σημαντικότεροι για έναν πράκτορα που ακολουθεί την επιβραβευμένη μάθηση.

4 Σχέση της επιβραβευμένης μάθησης με άλλα είδη μάθησης

Η επιβραβευμένη μάθηση διαφέρει σε αρκετά σημεία από την εποπτευόμενη μάθηση ή μάθηση υπό εποπτεία ή επίβλεψη (supervised learning). Η εποπτευόμενη μάθηση είναι μάθηση από παραδείγματα που παρέχονται από κάποιον εξωτερικό επιβλέποντα που έχει ήδη τις σχετικές γνώσεις. Αν και αυτό το είδος μάθησης είναι σημαντικό δεν αρκεί για να μάθει κανείς αλληλεπιδρώντας. Σε προβλήματα αλληλεπίδρασης είναι συχνά μη πρακτικό να εξασφαλίσει κανείς παραδείγματα επιθυμητής συμπεριφοράς που να είναι σωστά και αντιπροσωπευτικά όλων των καταστάσεων στις οποίες πρέπει να δράσει ο πράκτορας. Σε άγνωστες περιοχές -στις οποίες θα περίμενε κανείς η μάθηση να είναι ιδιαίτερα αποδοτική- ένας πράκτορας θα πρέπει να είναι ικανός να μάθει από τη δική του εμπειρία.

Μία από τις προκλήσεις που αναδύονται στην επιβραβευμένη μάθηση και

όχι σε άλλα είδη μάθησης είναι η εξισορρόπηση μεταξύ εξερεύνησης και εκμετάλλευσης. Προκειμένου να επιτύχει ένα πράκτορας επιβραβευμένης μάθησης μεγάλη ανταμοιβή πρέπει να προτιμά πράξεις που έχει δοκιμάσει στο παρελθόν και τις έχει βρει αποτελεσματικές στην παραγωγή ανταμοιβής. Αλλά προκειμένου να ανακαλύψει τέτοιες πράξεις πρέπει να δοκιμάσει πράξεις που δεν έχει επιλέξει στο παρελθόν. Ο πράκτορας πρέπει άρα να εκμεταλλευθεί ό,τι ήδη γνωρίζει προκειμένου να επιτύχει ανταμοιβή, αλλά πρέπει και να εξερευνήσει προκειμένου να κάνει καλύτερες επιλογές πράξεων στο μέλλον. Το δίλημμα συνίσταται στο γεγονός ότι κάποιος δεν μπορεί να ακολουθήσει αποκλειστικά είτε την εκμετάλλευση είτε την εξερεύνηση χωρίς να αποτύχει. Ο πράκτορας πρέπει να δοκιμάσει μία ποικιλία πράξεων και προοδευτικά να προτιμήσει εκείνες που μοιάζουν να είναι οι καλύτερες. Σε ένα στοχαστικό έργο κάθε πράξη πρέπει να δοκιμασθεί πολλές φορές προκειμένου να εκτιμηθεί αξιόπιστα η αναμενόμενη ανταμοιβή της. Το θέμα της εξισορρόπησης εκμετάλλευσης και εξερεύνησης δεν προκύπτει καθόλου στην εποπτευόμενη μάθηση, όπου αυτή είναι συνήθως ήδη ορισμένη.

Άλλο ένα σημαντικό χαρακτηριστικό της επιβραβευμένης μάθησης είναι ότι λαμβάνει υπ' όψιν της ρητά όλο το πρόβλημα ενός πράκτορα κατευθυνόμενου προς ένα συγκεκριμένο στόχο που αλληλεπιδρά με ένα αβέβαιο περιβάλλον. Αυτό έρχεται σε αντίθεση με πολλές προσεγγίσεις που αφοσιώνονται σε υποπροβλήματα χωρίς να ασχολούνται με το πως αυτά ενσωματώνονται στο μεγαλύτερο πρόβλημα. Για παράδειγμα η μηχανική μάθηση μελετά την εποπτευόμενη μάθηση χωρίς να ορίζει ρητά πως μία τέτοια ικανότητα είναι τελικά χρήσιμη. Άλλοι ερευνητές αναπτύσσουν θεωρίες σχεδιασμού με γενικούς στόχους, αλλά χωρίς να λαμβάνουν υπ' όψιν το ρόλο του σχεδιασμού στη λήψη αποφάσεων πραγματικού χρόνου ή το ερώτημα από που θα προέλθουν τα προγνωστικά μοντέλα που είναι απαραίτητα για το σχεδιασμό. Παρόλα τα χρήσιμα αποτελέσματα που αυτές οι προσεγγίσεις αποφέρουν, η εστίασή τους σε μεμονωμένα υποπροβλήματα -σε αντίθεση με την επιβραβευμένη μάθηση- αποτελεί σημαντικό περιορισμό.

Η επιβραβευμένη μάθηση ακολουθεί τον αντίθετο δρόμο, αρχίζοντας με έναν πλήρη, αλληλεπιδραστικό πράκτορα που προσπαθεί να επιτύχει ένα στόχο. Όλοι οι πράκτορες επιβραβευμένης μάθησης έχουν ρητούς στόχους, μπορούν να ανιχνεύσουν όψεις του περιβάλλοντός τους και μπορούν να επιλέξουν πράξεις για να επηρεάσουν το περιβάλλον τους. Επιπλέον συνήθως υποτίθεται ότι ο πράκτορας πρέπει να λειτουργεί παρά το γεγονός ότι αντιμετωπίζει σημαντική αβεβαιότητα σχετικά με το περιβάλλον του. Όταν η επιβραβευμένη μάθηση περιλαμβάνει σχεδιασμό, πρέπει να αντιμετωπίσει την αλληλεπίδραση μεταξύ του σχεδιασμού και την επιλογή πράξεων σε πραγματικό χρόνο, καθώς και το

ερώτημα του πώς περιβαλλοντικά μοντέλα αποκτώνται και βελτιώνονται. Όταν η επιβραβευμένη μάθηση εμπλέκει και εποπτευόμενη μάθηση το κάνει για πολύ συγκεκριμένους λόγους που καθορίζουν ποιες δυνατότητες είναι κρίσιμες και ποιες όχι. Προκειμένου να υπάρχει πρόοδος στην έρευνα γύρω από τη μάθηση, σημαντικά υποπροβλήματα πρέπει να απομονώνονται και να μελετώνται, αλλά θα πρέπει να είναι υποπροβλήματα που ξεκινούν από σαφείς ρόλους από πλήρεις, αλληλεπιδραστικούς, πράκτορες που αναζητούν στόχους, ακόμη και αν δεν είναι γνωστές ήδη όλες οι λεπτομέρειες των πρακτόρων αυτών.

5 Παραδείγματα εφαρμογής της επιβραβευμένης μάθησης

Μπορεί κανείς να κατανοήσει καλύτερα την επιβραβευμένη μάθηση μελετώντας ορισμένα παραδείγματα και πιθανές εφαρμογές που έχουν καθοδηγήσει την ανάπτυξή της.

- Ένας προσαρμοστικός ελεγκτής ρυθμίζει τις παραμέτρους λειτουργίας ενός διυλιστηρίου πετρελαίου σε πραγματικό χρόνο. Ο ελεγκτής βελτιστοποιεί την εξισορρόπηση απόδοσης / κόστους / ποιότητας βασιζόμενος σε καθορισμένα οριακά κόστη χωρίς να ακολουθεί αυστηρά ένα σύνολο σημείων που προτάθηκαν αρχικά από ανθρώπους - μηχανικούς.
- Ένα κινούμενο robot αποφασίζει εάν θα εισέλθει σε ένα νέο δωμάτιο σε αναζήτηση επιπλέον απορριμάτων προς συλλογή ή εάν θα αρχίσει να προσπαθεί να βρει το δρόμο του προς το σταθμό επαναφόρτισης της μπαταρίας του. Λαμβάνει την απόφασή του βασιζόμενο στο πόσο γρήγορα και εύκολα ήταν σε θέση να βρει τον επαναφορτιστή του στο παρελθόν.
- Μία νεαρή γαζέλλα μετά βίας μπορεί να σταθεί στα πόδια της λίγα λεπτά αφότου γεννηθεί. Μισή ώρα αργότερα τρέχει με ταχύτητα 30 μίλια ανά ώρα.
- Η απόφαση ενός καλού σκακιστή για το πια κίνηση θα επιλέξει λαμβάνεται τόσο με σχεδιασμό -προσδοκία πιθανών απαντήσεων και ανταπαντήσεων- όσο και με άμεσες, διαισθητικές κρίσεις για το πόσο επιθυμητές είναι συγκεκριμένες θέσεις και κινήσεις.
- Ακόμη και η φαινομενικά απλή διαδικασία της προετοιμασίας του πρωινού φαγητού κάποιου περιλαμβάνει πολύπλοκες συσχετίσεις αντανακλαστικής

συμπεριφοράς και αλληλοσυνδεόμενων σχέσεων στόχων και υποστόχων. Τέτοιοι είναι η εύρεση του φαγητού, η απόκτησή του, η προετοιμασία του κοκ. Κάθε βήμα εμπλέκει μία σειρά κινήσεων των ματιών για να συλλεχθούν οι κατάλληλες πληροφορίες και να καθοδηγήσουν τις ανάλογες κινήσεις του σώματος και των μελών του. Ταχείες κρίσεις γίνονται συνεχώς σχετικά με το πως πρέπει λόγου χάρη να μεταφερθούν τα αντικείμενα. Κάθε βήμα καθοδηγείται από στόχους και εξυπηρετεί άλλους στόχους ακόμη υψηλότερου επιπέδου.

Τα παραδείγματα αυτά μοιράζονται χαρακτηριστικά που είναι τόσο θεμελιώδη που εύκολα παραβλέπονται. Όλα περιλαμβάνουν αλληλεπίδραση μεταξύ ενός ενεργού πράκτορα που λαμβάνει αποφάσεις και του περιβάλλοντός του, μέσα στο οποίο ο πράκτορας επιχειρεί να επιτύχει το στόχο του παρά την αβεβαιότητα σχετικά με το περιβάλλον του. Οι πράξεις του πράκτορα μπορούν να επηρεάσουν τη μελλοντική κατάσταση του περιβάλλοντός του, επηρεάζοντας έτσι και τις επιλογές και τις ευκαιρίες που του είναι διαθέσιμες αργότερα. Η σωστή επιλογή απαιτεί να λαμβάνονται υπ' όψιν έμμεσες και καθυστερημένες συνέπειες πράξεων και έτσι πιθανώς να απαιτεί πρόβλεψη ή και σχεδιασμό.

Παράλληλα σε όλα τα παραπάνω παραδείγματα τα αποτελέσματα των πράξεων δεν μπορούν να προβλεφθούν πλήρως και έτσι ο πράκτορας πρέπει να παρακολουθεί συχνά το περιβάλλον του και να αντιδρά ανάλογα. Όλα τα παραδείγματα περιλαμβάνουν στόχους που είναι σαφείς με την έννοια ότι ο πράκτορας μπορεί να κρίνει την πρόοδο προς το στόχο του στη βάση του τι μπορεί να ανιχνεύσει άμεσα. Για παράδειγμα ο ελεγκτής του διυλιστηρίου μπορεί να ελέγξει πόσο πετρέλαιο έχει παράξει.

Σε όλα αυτά τα παραδείγματα ο πράκτορας μπορεί να χρησιμοποιήσει την εμπειρία του για να βελτιώσει την απόδοσή του με τον καιρό. Για παράδειγμα ο σκακιστής βελτιώνει τη διαίσθησή του στην αξιολόγηση θέσεων και έτσι βελτιώνει τον τρόπο που παίζει. Η γνώση που ο πράκτορας φέρνει στο έργο του στην αρχή, είτε από προηγούμενη εμπειρία με σχετικά έργα, είτε επειδή είναι ενσωματωμένη σε αυτόν από το σχεδιασμό ή την εξέλιξη, επηρεάζει το τι είναι χρήσιμο ή εύκολο να μαθευτεί. Ωστόσο η αλληλεπίδραση με το περιβάλλον είναι απαραίτητη για τη ρύθμιση της συμπεριφοράς, προκειμένου να εκμεταλλευτεί ο πράκτορας συγκεκριμένα χαρακτηριστικά του έργου.

6 Βασικά στοιχεία ενός συστήματος επιβραβευμένης μάθησης

Πέρα από τον πράκτορα και το περιβάλλον, μπορεί κανείς να αναγνωρίσει τέσσερα βασικά υποστοιχεία ενός συστήματος επιβραβευμένης μάθησης: Μία τακτική (πολιτική) (policy), μία συνάρτηση ανταμοιβής (reward function), μία συνάρτηση αξίας (value function) και προαιρετικά ένα μοντέλο του περιβάλλοντος (model of the environment).

1. Η *τακτική* καθορίζει τον τρόπο που ο πράκτορας συμπεριφέρεται σε μία δεδομένη χρονική στιγμή. Γενικά η τακτική είναι μία αντιστοίχιση από αντιλαμβανόμενες καταστάσεις του περιβάλλοντος σε πράξεις που πρέπει να λάβουν χώρα σε αυτές τις καταστάσεις. Σε μερικές περιπτώσεις η τακτική μπορεί να είναι μία απλή συνάρτηση ή ένας πίνακας αντιστοιχίσεων, ενώ σε άλλες μπορεί να περιλαμβάνει εκτεταμένους υπολογισμούς, όπως σε μία διαδικασία αναζήτησης. Η τακτική είναι ο πυρήνας ενός πράκτορα επιβραβευμένης μάθησης, με την έννοια ότι από μόνη της είναι αρκετή για να καθορίσει τη συμπεριφορά. Γενικά οι τακτικές μπορεί να είναι στοχαστικές.
2. Η *συνάρτηση ανταμοιβής* καθορίζει το στόχο σε ένα πρόβλημα επιβραβευμένης μάθησης. Γενικά αντιστοιχεί αντιλαμβανόμενες καταστάσεις (ή ζεύγη καταστάσεων και σταθμών) του περιβάλλοντος σε ένα συγκεκριμένο αριθμό, την ανταμοιβή, δείχνοντας πόσο επιθυμητή είναι ουσιασικά η κατάσταση. Ο μοναδικός σκοπός ενός πράκτορα επιβραβευμένης μάθησης είναι να μεγιστοποιήσει τη συνολική ανταμοιβή που λαμβάνει μακροπρόθεσμα. Η συνάρτηση ανταμοιβής καθορίζει ποια είναι τα θετικά και ποια τα αρνητικά περιστατικά για έναν πράκτορα, με άλλα λόγια τα άμεσα και καθοριστικά χαρακτηριστικά του προβλήματος που αντιμετωπίζει ο πράκτορας. Ως τέτοια, η συνάρτηση ανταμοιβής πρέπει να είναι καθορισμένη και σταθερή. Μπορεί ωστόσο να χρησιμοποιηθεί ως βάση για αλλαγή τακτικής. Για παράδειγμα, εάν μία πράξη που επιλέχθηκε από μία συγκεκριμένη τακτική έχει χαμηλή ανταμοιβή, τότε η τακτική μπορεί να αλλαχθεί για να επιλέξει μία άλλη πράξη σε αυτήν την κατάσταση στο μέλλον. Γενικά οι συναρτήσεις ανταμοιβής μπορεί επίσης να είναι στοχαστικές.
3. Ενώ η συνάρτηση ανταμοιβής δείχνει τι είναι καλό με μία άμεση έννοια, η *συνάρτηση αξίας* καθορίζει τι είναι καλό μακροπρόθεσμα. Γενικά η

αξία μίας κατάστασης είναι το σύνολο των ανταμοιβών που ένας πράκτορας μπορεί να περιμένει να συσσωρεύσει στο μέλλον αρχίζοντας από την κατάσταση αυτή. Ενώ οι ανταμοιβές καθορίζουν την άμεση, ουσιαστική επιθυμητότητα μίας κατάστασης του περιβάλλοντος, οι αξίες δείχνουν τη μακροπρόθεσμη επιθυμητότητα των καταστάσεων αφού ληφθούν υπ' όψιν οι καταστάσεις που είναι πιθανό να ακολουθήσουν και οι ανταμοιβές που θα είναι διαθέσιμες στις καταστάσεις αυτές. Για παράδειγμα, μία κατάσταση μπορεί να αποφέρει πάντα μία χαμηλή άμεση ανταμοιβή, αλλά να έχει ακόμη υψηλή αξία επειδή ακολουθείται τακτικά από άλλες καταστάσεις που αποφέρουν μεγάλες ανταμοιβές. Το αντίστροφο μπορεί επίσης να συμβαίνει. Κάτι αντίστοιχο συμβαίνει και με τους ανθρώπους, όπου οι ανταμοιβές είναι σαν την απόλαυση (εάν είναι υψηλές) ή σαν τον πόνο (εάν είναι χαμηλές), ενώ οι αξίες αντιστοιχούν σε μία πιο εκλεπτυσμένη και διορατική κρίση του πόσο ευχαριστημένος ή δυσαρεστημένος είναι κανείς όταν το περιβάλλον του είναι σε μία συγκεκριμένη κατάσταση.

Οι ανταμοιβές είναι κατά κάποιον τρόπο πρωτεύουσες, ενώ οι αξίες -ως προβλέψεις ανταμοιβών- είναι δευτερεύουσες. Χωρίς ανταμοιβές δε θα μπορούσαν να υπάρξουν αξίες και ο μόνος σκοπός του υπολογισμού αξιών είναι η επίτευξη μεγαλύτερων ανταμοιβών. Παρ' όλα αυτά οι αξίες είναι αυτές που μας απασχολούν περισσότερο όταν λαμβάνουμε ή αξιολογούμε αποφάσεις. Οι επιλογές πράξεων γίνονται στη βάση των κρίσεων περί αξιών. Αναζητούμε πράξεις που επιφέρουν καταστάσεις μεγαλύτερης αξίας, όχι μεγαλύτερης ανταμοιβής, διότι οι πράξεις αυτές φέρνουν το μεγαλύτερο ποσό ανταμοιβής μακροπρόθεσμα. Στη λήψη αποφάσεων και στο σχεδιασμό η αποκομιζόμενη ποσότητα που καλείται αξία είναι αυτή με την οποία ασχολούμαστε περισσότερο. Δυστυχώς είναι και πολύ πιο δύσκολος ο καθορισμός αξιών από τον καθορισμό ανταμοιβών. Οι ανταμοιβές δίνονται βασικά άμεσα από το περιβάλλον, ενώ οι αξίες πρέπει να εκτιμηθούν και να επανεκτιμηθούν από τη διαδοχή των παρατηρήσεων που κάνει ένας πράκτορας καθ' όλη τη διάρκεια ζωής του. Στην πραγματικότητα, το πιο σημαντικό συστατικό σχεδόν όλων των αλγορίθμων επιβραβευμένης μάθησης είναι η μέθοδος αποτελεσματικής εκτίμησης αξιών.

Παρόλο που οι περισσότερες μέθοδοι επιβραβευμένης μάθησης δομούνται γύρω από συναρτήσεις εκτίμησης αξίας, αυτό δεν είναι εντελώς απαραίτητο προκειμένου να λυθούν προβλήματα επιβραβευμένης μάθησης. Για παράδειγμα μέθοδοι αναζήτησης όπως οι γενετικοί αλγόριθμοι, ο γενετικός προγραμματισμός, η προσομοιωμένη απόκτηση και άλλες μέθοδοι βελτιστοποίησης συναρτήσεων έχουν χρησιμοποιηθεί για να λύσουν προβλήματα επιβραβευμένης μάθησης. Οι μέθοδοι αυτές αναζητούν άμεσα

στο χώρο των τακτικών, χωρίς να επικαλούνται συναρτήσεις αξίας. Οι μέθοδοι αυτές ονομάζονται εξελικτικές, γιατί η λειτουργία τους είναι ανάλογη του τρόπου που η βιολογική εξέλιξη παράγει οργανισμούς με έμπειρη συμπεριφορά, παρόλο που οι ίδιοι δεν μαθαίνουν στη διάρκεια της ζωής τους. Εάν ο χώρος των τακτικών είναι επαρκώς μικρός, ή μπορεί να δομηθεί έτσι ώστε οι καλές τακτικές να είναι συνήθεις και εύκολα εντοπίσιμες, τότε συχνά οι εξελικτικές μέθοδοι είναι αποτελεσματικές. Επιπρόσθετα οι εξελικτικές μέθοδοι πλεονεκτούν σε προβλήματα στα οποία ο πράκτορας δεν μπορεί να ανιχνεύσει με ακρίβεια την κατάσταση του περιβάλλοντός του.

Εντούτοις, η επιβραβευμένη μάθηση περιλαμβάνει μάθηση κατά τη διάρκεια της αλληλεπίδρασης με το περιβάλλον, κάτι το οποίο οι εξελικτικές μέθοδοι δεν κάνουν. Μέθοδοι όμως που λαμβάνουν υπ' όψιν και εκμεταλλεύονται τις λεπτομέρειες των συγκεκριμένων κάθε φορά αλληλεπιδράσεων συμπεριφοράς μπορούν να είναι πιο αποτελεσματικές από τις εξελικτικές μεθόδους σε πολλές περιπτώσεις. Οι εξελικτικές μέθοδοι αγνοούν μεγάλο μέρος της χρήσιμης δομής των προβλημάτων επιβραβευμένης μάθησης, μιας και δεν χρησιμοποιούν το γεγονός ότι η τακτική που αναζητούν είναι μία συνάρτηση από καταστάσεις σε πράξεις και δεν αντιλαμβάνονται ποιες καταστάσεις πρέπει να περάσει κανείς κατά τη διάρκεια ζωής του ή ποιες πράξεις επιλέγει. Σε μερικές περιπτώσεις η πληροφορία αυτή μπορεί να είναι παραπλανητική, για παράδειγμα όταν οι καταστάσεις δεν έχουν γίνει σωστά αντιληπτές, αλλά συνήθως επιτρέπει πιο αποτελεσματική αναζήτηση, μιας και αντλούνται συμπεράσματα καθ' όλη τη διάρκεια της διαδικασίας και όχι μόνο από το τελικό αποτέλεσμα. Παρά το ότι η εξέλιξη και η μάθηση μοιράζονται ορισμένα χαρακτηριστικά και μπορούν να συνεργασθούν φυσικά, δε θα έλεγε κανείς πως οι εξελικτικές μέθοδοι είναι ιδιαίτερα κατάλληλες για την επίλυση προβλημάτων επιβραβευμένης μάθησης.

4. Το τέταρτο και τελευταίο στοιχείο μερικών συστημάτων επιβραβευμένης μάθησης είναι ένα μοντέλο του περιβάλλοντος, το οποίο ουσιαστικά μιμείται τη συμπεριφορά του περιβάλλοντος. Για παράδειγμα για μια δεδομένη κατάσταση και πράξη το μοντέλο μπορεί να προβλέψει την επακόλουθη επόμενη κατάσταση και ανταμοιβή. Τα μοντέλα χρησιμοποιούνται για το σχεδιασμό, εννοώντας κάθε τρόπο απόφασης σε μια σειρά πράξεων, λαμβάνοντας υπ' όψιν πιθανές μελλοντικές καταστάσεις πριν αυτές πραγματοποιηθούν. Η ενσωμάτωση μοντέλων και σχεδιασμού σε συστήματα επιβραβευμένης μάθησης είναι μία σχετικά πρόσφατη εξέλιξη. Τα πρώτα συστήματα επιβραβευμένης μάθησης ήταν σαφώς μαθητές δοκιμής και λάθους, με άλλα λόγια δε σχεδίαζαν καθόλου. Ωστόσο έγινε σταδιακά

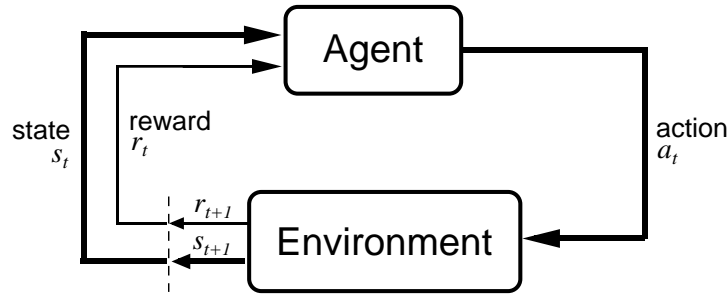
εμφανές ότι οι μέθοδοι επιβραβευμένης μάθησης συνδέονται στενά με τις μεθόδους δυναμικού προγραμματισμού, οι οποίες χρησιμοποιούν μοντέλα και οι οποίες με τη σειρά τους είναι στενά συνδεδεμένες με μεθόδους σχεδιασμού καταστάσεων και χώρου. Η σύγχρονη επιβραβευμένη μάθηση καλύπτει όλο το φάσμα από τη χαμηλού επιπέδου μάθηση δοκιμής και λάθους μέχρι τις υψηλού επιπέδου μεθόδους μάθησης που χρησιμοποιούν μοντέλα του περιβάλλοντος ή και σχεδιασμό.

7 Βασικές κατηγορίες μεθόδων επιβραβευμένης μάθησης

Ίσως το σημαντικότερο χαρακτηριστικό της επιβραβευμένης μάθησης που τη διαφοροποιεί από άλλα είδη μάθησης είναι ότι χρησιμοποιεί πληροφορία εκπαίδευσης που αξιολογεί τις πράξεις που έχουν γίνει, αντί να καθοδηγεί δίνοντας τη σωστή πράξη. Αυτό δημιουργεί την ανάγκη για ενεργή εξερεύνηση, για αναζήτηση της καλής συμπεριφοράς με τη μέθοδο δοκιμής και λάθους. Καθαρά αξιολογητικές μέθοδοι, όπως η βελτιστοποίηση συναρτήσεων, που χρησιμοποιούν και οι εξελικτικές μέθοδοι, δείχνουν πόσο καλή ήταν η πράξη που έγινε κάθε φορά, αλλά όχι αν ήταν η καλύτερη ή η χειρότερη δυνατή πράξη. Από την άλλη πλευρά καθαρά καθοδηγητικές μέθοδοι, όπως η εποπτευόμενη μάθηση, που χρησιμοποιείται μεταξύ άλλων στην αναγνώριση προτύπων και στα τεχνητά νευρωνικά δίκτυα, δείχνει τη σωστή πράξη που πρέπει να γίνει, ανεξάρτητα από την πράξη που τελικά έγινε στην πραγματικότητα. Τα παραπάνω δείχνουν τη βασική διαφορά των δύο τρόπων ανάδρασης: Η αξιολογητική ανάδραση εξαρτάται πλήρως από την πράξη που τελικά έγινε, ενώ η καθοδηγητική ανάδραση είναι ανεξάρτητη αυτής. Υπάρχουν βέβαια και περιπτώσεις μείξης των δύο μεθόδων.

Παράλληλα υπάρχουν μέθοδοι συσχετιστικές, σύμφωνα με τις οποίες οι είσοδοι αντιστοιχούνται σε εξόδους και μαθαίνεται η καλύτερη έξοδος για κάθε είσοδο και μέθοδοι μη συσχετιστικές, σύμφωνα με τις οποίες “ μαθαίνεται ”, δηλαδή βρίσκεται μία -η γενικά καλύτερη- έξοδος.

Γενικά υπάρχουν διάφοροι απλοί τρόποι εξισορρόπησης μεταξύ της εξερεύνησης νέων πράξεων και της εκμετάλλευσης των ήδη γνωστών. Οι greedy μέθοδοι διαλέγουν τυχαία ένα μικρό ποσοστό του χρόνου, οι softmax μέθοδοι διαβαθμίζουν τις πιθανότητες των πράξεών τους σύμφωνα με τις τρέχουσες εκτιμήσεις πράξεων και αξιών και οι pursuit μέθοδοι απλώς εξακολουθούν να βαδίζουν προς την τρέχουσα “ άπληστη ” πράξη.



Σχήμα 1: Αλληλεπίδραση του πράκτορα επιβραβευμένης μάθησης και του περιβάλλοντός του.

8 Βασικά στοιχεία του προβλήματος της επιβραβευμένης μάθησης

Η επιβραβευμένη μάθηση αφορά το πως κανείς μαθαίνει από την αλληλεπίδραση πως να συμπεριφέρεται προκειμένου να επιτύχει ένα στόχο. Ο πράκτορας (agent) επιβραβευμένης μάθησης και το περιβάλλον (environment) του αλληλεπιδρούν όπως δείχνει και το σχήμα 1 κατά τη διάρκεια μιας ακολουθίας διακριτών χρονικών βημάτων. Οι πράξεις (actions) είναι οι επιλογές που κάνει ο πράκτορας, οι καταστάσεις (states) είναι η βάση για να κάνει αυτός επιλογές και οι ανταμοιβές (rewards) είναι η βάση για την αξιολόγηση των επιλογών. Κάθετι εντός του πράκτορα είναι πλήρως γνωστό και ελέγξιμο από αυτόν, ενώ κάθετι εκτός αυτού δεν είναι πλήρως ελέγξιμο και είναι κατά περίπτωση πλήρως γνωστό. Μία τακτική (policy) είναι ένας στοχαστικός κανόνας σύμφωνα με τον οποίο ο πράκτορας επιλέγει πράξεις ως συνάρτηση καταστάσεων. Ο σκοπός του πράκτορα είναι να μεγιστοποιήσει το ποσό ανταμοιβής που λαμβάνει με το χρόνο.

Η ανταπόδοση (return) είναι η συνάρτηση των μελλοντικών ανταμοιβών, την οποία ο πράκτορας προσπαθεί να μεγιστοποιήσει. Οι ορισμοί της διαφέρουν, ανάλογα με το αν ενδιαφέρεται κανείς για τη συνολική (total) ή για κάποια μειωμένη (discounted) ανταμοιβή. Η πρώτη είναι κατάλληλη για εργασίες που χωρίζονται σε επεισόδια, στις οποίες η αλληλεπίδραση πράκτορα και περιβάλλοντος χωρίζεται φυσικά σε επεισόδια, ενώ η δεύτερη είναι κατάλληλη για συνεχείς εργασίες, στις οποίες η αλληλεπίδραση δε χωρίζεται φυσικά σε επεισόδια, αλλά συνεχίζεται απεριόριστα.

Ένα περιβάλλον ικανοποιεί την ιδιότητα Markov (Markov property) εάν η κατάσταση του συμπυκνώνει το παρελθόν χωρίς να υποβαθμίζει την ικανότητα

της πρόβλεψης του μέλλοντος. Αυτό πολύ σπάνια είναι απόλυτα αληθινό, αλλά συχνά το σήμα της κατάστασης μπορεί να επιλεγεί ή να κατασκευασθεί έτσι, ώστε η ιδιότητα αυτή να ισχύει προσεγγιστικά. Στις απλές περιπτώσεις που εξετάζουμε υποθέτουμε ότι αυτό έχει ήδη γίνει και εστιάζουμε την προσοχή μας στο πρόβλημα της λήψης της εξής απόφασης: Πώς δηλαδή θα αποφασίσουμε τι θα κάνουμε, συναρτήσει οποιουδήποτε σήματος κατάστασης είναι διαθέσιμο. Εάν η ιδιότητα Markov ισχύει, τότε το περιβάλλον ονομάζεται Διαδικασία Απόφασης Markov (Markov Decision Process - MDP). Μία πεπερασμένη (finite) MDP είναι μία MDP με πεπερασμένα σύνολα καταστάσεων και πράξεων και είναι αυτή η περίπτωση που συνήθως μας απασχολεί. Στην περίπτωση που ο πράκτορας μπορεί να δει μόνο ένα τμήμα της κατάστασης του περιβάλλοντός του έχει την ανάγκη εσωτερικής μνήμης καταστάσεων για να δρα βέλτιστα και μιλάμε για μοντέλο Μερικά Παρατηρήσιμης MDP (Partially Observable Markov Decision Process - POMDP).

Η συνάρτηση αξίας (value function) μίας τακτικής αντιστοιχεί σε κάθε κατάσταση την αναμενόμενη ανταπόδοση από την κατάσταση αυτή, δεδομένου ότι ο πράκτορας θα χρησιμοποιήσει την τακτική αυτή. Η βέλτιστη συνάρτηση αξίας (optimal value function) αντιστοιχεί σε κάθε κατάσταση τη μέγιστη αναμενόμενη ανταπόδοση που μπορεί να επιτευχθεί από οποιαδήποτε τακτική. Μια τακτική της οποίας η συνάρτηση αξίας είναι η βέλτιστη συνάρτηση αξίας είναι μία βέλτιστη τακτική (optimal policy). Ενώ για κάθε MDP υπάρχει μόνο μία βέλτιστη συνάρτηση αξίας, μπορούν να υπάρχουν πολλές βέλτιστες τακτικές. Κάθε τακτική που είναι άπληστη σε σχέση με τη βέλτιστη συνάρτηση αξίας είναι βέλτιστη τακτική. Η εξίσωση ιδανικότητας του Bellman (Bellman optimality equation) είναι μία ειδική συνθήκη συνέπειας που πρέπει να ικανοποιεί η βέλτιστη συνάρτηση αξίας και που μπορεί να λυθεί για αυτήν και από την οποία μπορεί να καθορισθεί σχετικά εύκολα μία βέλτιστη τακτική.

Ένα πρόβλημα επιβραβευμένης μάθησης μπορεί να διατυπωθεί με διάφορους τρόπους, ανάλογα με τις υποθέσεις που γίνονται σχετικά με το επίπεδο της γνώσης που είναι αρχικά διαθέσιμο στον πράκτορα. Σε προβλήματα πλήρους γνώσης ο πράκτορας έχει ένα πλήρες και ακριβές μοντέλο της δυναμικής του περιβάλλοντός του. Εάν το περιβάλλον είναι MDP, τότε κάθε τέτοιο μοντέλο αποτελείται από τις πιθανότητες μετάβασης ενός βήματος και τις αναμενόμενες ανταμοιβές για όλες τις καταστάσεις και τις αντίστοιχες επιτρεπτές πράξεις. Σε προβλήματα μη πλήρους γνώσης ένα πλήρες και τέλειο μοντέλο του περιβάλλοντος δεν είναι διαθέσιμο.

Ακόμη και αν ο πράκτορας έχει στη διάθεσή του ένα πλήρες και ακριβές μοντέλο του περιβάλλοντος, συνήθως δεν είναι σε θέση να εκτελέσει αρκετούς

υπολογισμούς σε κάθε βήμα για να το χρησιμοποιήσει πλήρως. Η διαθέσιμη μνήμη είναι συχνά ένας σημαντικός περιορισμός, μιας και αυτή χρησιμοποιείται για να κατασκευασθούν ακριβείς προσεγγίσεις των συναρτήσεων αξίας, των τακτικών και των μοντέλων. Στις περισσότερες περιπτώσεις πρακτικού ενδιαφέροντος υπάρχουν πολύ περισσότερες καταστάσεις από όσες θα μπορούσαν πιθανά να αποθηκευθούν σε πίνακες και οι προσεγγίσεις είναι απαραίτητες. Την επιβραβευμένη μάθηση απασχολούν ιδιαίτερα περιπτώσεις στις οποίες δεν μπορεί να βρεθεί βέλτιστη λύση και πρέπει να γίνουν προσεγγίσεις.

9 Θεμελιώδεις μέθοδοι της επιβραβευμένης μάθησης

Στη συνέχεια παρουσιάζονται τρεις θεμελιώδεις κλάσεις μεθόδων για τη λύση του προβλήματος της επιβραβευμένης μάθησης: Ο δυναμικός προγραμματισμός (Dynamic Programming - DP), οι μέθοδοι Monte Carlo (MC methods) και η μέθοδος μάθησης με χρονική διαφορά (temporal-difference learning). Καθεμιά από αυτές τις τρεις κλάσεις μεθόδων έχει τα πλεονεκτήματά της και τις αδυναμίες της. Οι μέθοδοι δυναμικού προγραμματισμού είναι πολύ καλά ανεπτυγμένες μαθηματικά, αλλά απαιτούν ένα πλήρες και ακριβές μοντέλο του περιβάλλοντος. Οι μέθοδοι Monte Carlo δεν απαιτούν κάποιο μοντέλο, αλλά ενώ είναι απλή η βασική τους ιδέα δεν ταιριάζουν σε υπολογισμούς που πρέπει να γίνουν βήμα προς βήμα. Τέλος, οι μέθοδοι χρονικής διαφοράς δεν απαιτούν κάποιο μοντέλο, αλλά είναι περισσότερο πολύκλοκες στην ανάλυση. Υπάρχουν ακόμη διαφορές μεταξύ των μεθόδων σε σχέση με την αποδοτικότητά τους και την ταχύτητα σύγκλισης.

10 Δυναμικός Προγραμματισμός

Ο όρος “ Δυναμικός Προγραμματισμός ” (*Dynamic Programming - DP*) αναφέρεται σε μια συλλογή αλγορίθμων, που μπορούν να χρησιμοποιηθούν για τον υπολογισμό της βέλτιστης τακτικής ώστε να πάρουμε το τέλειο μοντέλο από το περιβάλλον με τη μορφή μιας Markov διαδικασίας απόφασης (Markov Decision Process - MDP). Οι κλασικοί αλγόριθμοι δυναμικού προγραμματισμού χρησιμοποιούνται ελάχιστα στην επιβραβευμένη μάθηση λόγω της μεγάλης υπολογιστικής έκτασης και των προϋποθέσεων για ένα τέλειο μοντέλο. Όμως θεωρητικά είναι ακόμη πολύ σημαντικοί. Ο δυναμικός προγραμματισμός

ωστόσο είναι θεμελιώδης για την κατανόηση των υπολοίπων μεθόδων. Στην πραγματικότητα άλλωστε όλες οι άλλες μέθοδοι μπορούν να θεωρηθούν προσπάθειες για την επίτευξη του ίδιου αποτελέσματος με αυτό του δυναμικού προγραμματισμού, αλλά με λιγότερους υπολογισμούς και χωρίς τη χρήση ενός ιδανικού μοντέλου. Ξεκινώντας, θεωρούμε το περιβάλλον σαν μια πεπερασμένη MDP, η οποία αποτελείται από:

- Ένα σύνολο καταστάσεων (states) S .
- Ένα σύνολο πράξεων (actions) A .
- Μία συνάρτηση ανταμοιβής (reward function) $R : S \times A \rightarrow \mathbb{R}$.
- Μία συνάρτηση μετάβασης κατάστασης (state transition function) $T : S \times A \rightarrow P(S)$, όπου $P(S)$ είναι η πιθανότητα κατανομής στο σύνολο S . Λέμε πως $T(s, a, s')$ είναι η πιθανότητα μετάβασης από την κατάσταση s στην κατάσταση s' χρησιμοποιώντας τη λειτουργία a .

Παρόλο που η ιδέα του δυναμικού προγραμματισμού εφαρμόζεται σε προβλήματα συνεχών διαστημάτων κατάστασης και λειτουργίας, ακριβείς λύσεις είναι δυνατές μόνο σε ειδικές περιπτώσεις. Ένας κοινός τρόπος για να λαμβάνουμε προσεγγιστικές λύσεις είναι η κβάντιση των διαστημάτων κατάστασης και λειτουργίας και ακολούθως η εφαρμογή finite-state DP μεθόδων. Το κλειδί στον δυναμικό προγραμματισμό, και στην επιβραβευμένη μάθηση γενικότερα, είναι η χρησιμοποίηση συναρτήσεων εκτίμησης (value functions) ώστε να οργανωθεί αλλά και να δομηθεί η έρευνα καλών τακτικών.

10.1 Αξιολόγηση τακτικής (policy evaluation)

Αρχικά θα μελετήσουμε τον υπολογισμό της συνάρτησης state-value, V^* , για μια αυθαίρετη τακτική π . Ονομάζεται αξιολόγηση τακτικής (policy evaluation) και συχνά αναφέρεται και ως πρόβλημα πρόβλεψης (prediction problem). Θεωρώντας την πλήρη απόφαση τακτικής, γράφουμε:

$$V^*(s) = \max_{\pi} E(\sum_{t=0}^{\infty} \gamma^t r_t)$$

Αυτή η ιδανική *value function* είναι μοναδική και μπορεί να οριστεί ως η λύση ταυτόχρονων εξισώσεων:

$$V^*(s) = \max_a (R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V^*(s')), \forall s \in S, \quad (1)$$

που υποστηρίζουν πως η τιμή / αξία της κατάστασης s είναι η αναμενόμενη ακαριαία (αντ)αμοιβή και η αναμενόμενη μειωμένη τιμή της επόμενης κατάστασης, χρησιμοποιώντας την καλύτερη διαθέσιμη λειτουργία.

Έχοντας τώρα την ιδανική value function μπορούμε να προσδιορίσουμε την ιδανική policy:

$$\pi^*(s) = \operatorname{argmax}_a (R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V^*(s'))$$

1. Αλγόριθμος επαναλήψεων αξίας (value iteration). Ένας τρόπος για να βρεθεί η ιδανική policy είναι να βρεθεί πρώτα η ιδανική value function. Κάτι τέτοιο είναι δυνατό μέσω ενός επαναληπτικού αλγορίθμου, που ονομάζεται value iteration:

```

initialize  $V(s)$  arbitrarily
loop until policy good enough
  loop for  $s \in S$ 
    loop for  $a \in A$ 
       $Q(s, a) := R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V(s')$ 
       $V(s) := \max_a Q(s, a)$ 
    end loop
  end loop
end loop

```

Δεν είναι ξεκάθαρο το πότε σταματά ο αλγόριθμος αυτός. Ένα, όμως, σημαντικό αποτέλεσμα περιορίζει την εκτέλεση της τρέχουσας “greedy policy”, ως μια συνάρτηση του υπολοίπου Bellman της τρέχουσας τιμής της συνάρτησης. Σύμφωνα λοιπόν με αυτό, αν η μέγιστη διαφορά μεταξύ δύο επιτυχών τιμών συνάρτησης είναι μικρότερη του ϵ , τότε η τιμή της greedy policy, (greedy policy (άπληστη τακτική) είναι η τακτική που προκύπτει αν διαλέξουμε, σε κάθε κατάσταση, τη λειτουργία που μεγιστοποιεί την εκτιμώμενη προεξοφλούμενη ανταμοιβή) διαφέρει από την τιμή της συνάρτησης της ιδανικής policy όχι περισσότερο από $2\epsilon\gamma/(1-\gamma)$ σε κάθε κατάσταση. Για να σταματήσει, επομένως, ο αλγόριθμος το κριτήριο αυτό είναι ιδιαίτερα αποτελεσματικό.

Αξίζει να σημειωθεί ότι ο αλγόριθμος value iteration είναι πολύ ευέλικτος. Οι υπολογισμοί του V δεν είναι απαραίτητο να γίνονται με αυστηρή σειρά. Αντίθετα, μπορούν να συμβαίνουν ασύγχρονα και παράλληλα,

ώστε η τιμή της κάθε κατάστασης να ανανεώνεται συνεχώς σε μία συνεχή εκτέλεση.

Οι ανανεώσεις (updates) που βασίζονται στην εξίσωση (1) είναι γνωστές ως full backups, γιατί κάνουν χρήση πληροφορίας από όλες τις δυνατές διάδοχες καταστάσεις. Ανανεώσεις της μορφής :

$$Q(s, a) := Q(s, a) + a(r + \gamma \max_{a'} Q(s', a') - Q(s, a))$$

μπορούν να συμβαίνουν όσο κάθε ζευγάρι των a και s ανανεώνεται απείρως συχνά, το s' δειγματοληπτείται από την κατανομή $T(s, a, s')$, το r δειγματοληπτείται με μέση τιμή $R(s, a)$ και περιορισμένη διακύμανση, και ο ρυθμός μάθησης μειώνεται αργά.

Η υπολογιστική πολυπλοκότητα του αλγορίθμου value iteration με full backups, ανά επανάληψη, είναι δευτέρου βαθμού στον αριθμό των καταστάσεων και γραμμική στον αριθμό των λειτουργιών. Συνήθως, οι πιθανότητες μετάβασης $T(s, a, s')$ είναι σποραδικές. Αν υπάρχει κατά μέσο όρο ένας σταθερός αριθμός επόμενων καταστάσεων με μη μηδενική πιθανότητα τότε το κόστος ανά επανάληψη είναι γραμμικό όσον αφορά τον αριθμό των καταστάσεων και γραμμικό όσον αφορά τον αριθμό των λειτουργιών. Ο αριθμός των επαναλήψεων που απαιτούνται για να φτάσουμε στην ιδανική τιμή συνάρτησης είναι πολυώνυμο στον αριθμό των καταστάσεων και στο μέγεθος της μεγαλύτερης ανταμοιβής, αν ο μειούμενος παράγοντας διατηρείται σταθερός. Στην χειρότερη περίπτωση ο αριθμός των επαναλήψεων αυξάνει πολυωνυμικά σε $1/(1-\gamma)$, οπότε ο ρυθμός σύγκλισης καθυστερεί όσο ο μειούμενος παράγοντας πλησιάζει τη μονάδα.

2. Αλγόριθμος επαναλήψεων τακτικής (policy iteration).

Ο αλγόριθμος policy iteration παράγει την policy απευθείας, και όχι έμμεσα, μέσω της ιδανικής value function. Λειτουργεί ως εξής:

choose an arbitrary policy π'

loop

$$\pi := \pi'$$

compute the value function of policy π :

solve the linear equations

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} T(s, \pi(s), s') V_\pi(s')$$

improve the policy at each state:

$$\pi'(s) := \operatorname{argmax}_a (R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V_\pi(s'))$$

until $\pi = \pi'$

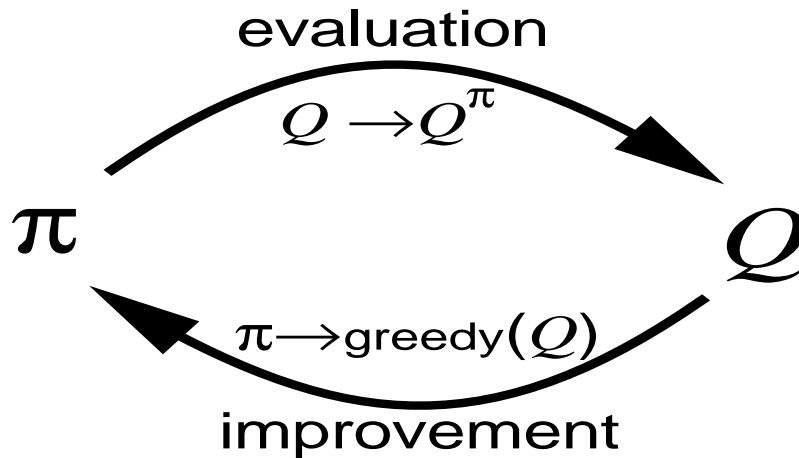
Η value function μιας policy είναι ακριβώς η αναμενόμενη άπειρη προεξοφλούμενη ανταμοιβή (discounted reward), που θα κερδίσουμε σε κάθε κατάσταση, αν εκτελέσουμε αυτή την policy. Κάτι τέτοιο επιτυγχάνεται λύνοντας ένα σύνολο γραμμικών εξισώσεων. Αν γνωρίζουμε την τιμή κάθε κατάστασης υπό την τρέχουσα policy, εξετάζουμε αν η τιμή αυτή θα μπορούσε να βελτιωθεί αλλάζοντας την πρώτη λειτουργία. Αν γίνεται, αλλάζουμε την policy για να πάρουμε τη νέα λειτουργία, όποτε είναι σ' αυτή την θέση. Αυτό το βήμα εγγυάται αυστηρά βελτίωση της εκτέλεσης της policy. Όταν οι βελτιώσεις δεν είναι δυνατές, τότε η policy είναι η ιδανική.

Από τη στιγμή που υπάρχουν το πολύ $|A|^{|S|}$ ευδιάκριτες policies και η διαδοχή τους βελτιώνεται σε κάθε βήμα, ο αλγόριθμος αυτός τελειώνει το πολύ σε έναν εκθετικό αριθμό από επαναλήψεις. Ωστόσο, είναι ένα καλό ερώτημα πόσες επαναλήψεις χρειάζονται στην χειρότερη περίπτωση. Είναι γνωστό πως ο χρόνος εκτέλεσης είναι ένα ψευδοπολυώνυμο και για κάθε προκαθορισμένο παράγοντα μείωσης υπάρχει ένα όριο, πολυώνυμο για το συνολικό μέγεθος μιας MDP.

10.2 Γενικευμένη επανάληψη τακτικών (generalized policy iteration)

Ο αλγόριθμος policy iteration περιλαμβάνει δύο ταυτόχρονες διεργασίες, που αλληλοεπιδρούν. Η μια υπολογίζει την value function σύμφωνα με την τρέχουσα policy (policy evaluation) και η άλλη υπολογίζει την greedy policy σύμφωνα με την τρέχουσα value function (policy improvement). Στην περίπτωση του policy iteration αυτές οι δύο διεργασίες εναλλάσσονται, μόλις δηλαδή ολοκληρώνεται η μία ξεκινά η άλλη. Όμως αυτό δεν είναι και απαραίτητο. Στην περίπτωση του value iteration, για παράδειγμα, μόνο μια επανάληψη της policy evaluation διεξάγεται σε κάθε διεργασία policy improvement. Στις ασύγχρονες μεθόδους δυναμικού προγραμματισμού, οι διεργασίες evaluation και improvement παρεμβάλλονται μεταξύ τους πολύ πιο συχνά. Σε μερικές περιπτώσεις μια μόνο κατάσταση είναι ανανεωμένη σε μια διεργασία πριν τρέξει στην άλλη. Όσο και οι δυο διεργασίες συνεχίζουν να ανανεώνουν όλες τις καταστάσεις, το τελικό αποτέλεσμα είναι τυπικά το ίδιο -σύγκλιση στην optimal value function και μια optimal policy.

Χρησιμοποιείται ο όρος generalized policy iteration (GPI) όταν απευθυνόμαστε στην γενική ιδέα του να αφήνουμε τις διεργασίες policy evaluation και policy improvement να εναλλάσσονται, όπως παρουσιάζεται στο σχήμα 2.



Σχήμα 2: Generalized policy iteration.

Σχεδόν όλες οι μέθοδοι της επιβραβευμένης μάθησης περιγράφονται ως GPI. Αυτό σημαίνει πως όλες έχουν ξεχωριστές policies και value functions, με την policy συνεχώς να βελτιώνεται σε σχέση με την value function και τη value function να οδηγείται πάντα προς τη value function για την policy.

Οι διεργασίες evaluation και improvement στην GPI μπορεί να θεωρηθούν ότι συναγωνίζονται και συνεργάζονται ταυτόχρονα. Συναγωνίζονται με την έννοια ότι τραβούν προς αντίθετες κατευθύνσεις. Κάνοντας την policy greedy σε σχέση με τη value function τυπικά η value function είναι λάθος για την αλλαγμένη policy. Και κάνοντας τη value function συνεπή με την policy προκύπτει η policy να μην είναι πια greedy. Βέβαια μετά από πολύ χρόνο, αυτές οι δυο συγκλίνουσες διεργασίες αλληλεπιδρούν για να βρουν μια μοναδική κοινή λύση: την optimal value function και την optimal policy.

10.3 Η αποδοτικότητα του δυναμικού προγραμματισμού

Ο Δυναμικός Προγραμματισμός μπορεί να μην είναι πρακτικός για πολύ μεγάλα προβλήματα, αλλά, συγκρινόμενος με άλλες μεθόδους λύσης Markov διεργασιών απόφασης (MDPs), είναι αρκετά αποδοτικός. Αν αγνοήσουμε κάποιες τεχνικές λεπτομέρειες, τότε ο χειρότερος χρόνος που απαιτείται για να βρεθεί η optimal policy είναι πολυώνυμο στον αριθμό των καταστάσεων και των πράξεων/ενεργειών. Αν n και m είναι αντίστοιχα ο αριθμός των καταστάσεων και των πράξεων, αυτό σημαίνει πως ο δυναμικός προγραμματισμός χρειάζεται έναν αριθμό υπολογιστικών ενεργειών, που είναι μικρότερος από κάποια πολυωνυμι-

κή συνάρτηση από n και m . Ο δυναμικός προγραμματισμός εγγυάται ότι θα βρει μια optimal policy σε πολυωνυμικό χρόνο ακόμα και αν ο συνολικός αριθμός (ντετερμινιστικών) policies είναι mn . Οπότε ο δυναμικός προγραμματισμός είναι εκθετικά ταχύτερος από οποιαδήποτε απευθείας έρευνα της policy, γιατί η απευθείας έρευνα θα έπρεπε να εξετάζει εξονυχιστικά κάθε policy για να υπάρχει η ίδια εγγύηση.

Ακόμη, γραμμικές μέθοδοι προγραμματισμού χρησιμοποιούνται για τη λύση Markov διαδικασιών απόφασης (MDPs) και σε μερικές περιπτώσεις οι εγγυήσεις στη χειρότερη περίπτωση σύγκλισης είναι καλύτερες από εκείνες των DP μεθόδων. Βέβαια οι γραμμικές μέθοδοι προγραμματισμού δεν είναι πρακτικές όταν υπάρχει μικρός αριθμός καταστάσεων. Για τα μεγάλα προβλήματα μόνο οι DP μέθοδοι είναι δυνατές.

Ο δυναμικός προγραμματισμός μερικές φορές θεωρείται περιορισμένης ικανότητας εφαρμογής λόγω της κατάρτας των διαστάσεων (curse of dimensionality). Αυτό σημαίνει ότι ο αριθμός των καταστάσεων συχνά αυξάνει εκθετικά με τον αριθμό των μεταβλητών κατάστασης. Μεγάλα σύνολα καταστάσεων προκαλούν δυσκολίες, όμως αυτές είναι έμφυτες δυσκολίες του προβλήματος, και όχι του δυναμικού προγραμματισμού ως μεθόδου λύσης.

Στην πράξη σήμερα, με τη βοήθεια των ηλεκτρονικών υπολογιστών οι μέθοδοι δυναμικού προγραμματισμού χρησιμοποιούνται για να λύσουν MDPs με εκατομμύρια καταστάσεις. Στα προβλήματα δε με μεγάλο αριθμό διαστημάτων κατάστασης προτιμούνται οι ασύγχρονες μέθοδοι δυναμικού προγραμματισμού.

11 Μέθοδοι Monte Carlo

Στο σημείο αυτό θα μελετήσουμε τις πρώτες μεθόδους μάθησης για την εκτίμηση των value functions και την ανακάλυψη των optimal policies. Σε αντίθεση με την προηγούμενη μέθοδο, εδώ δεν κατέχουμε πλήρη γνώση του περιβάλλοντος. Οι έμπημέθοδοι Monte Carlo απαιτούν μόνο εμπειρία. Δειγματοληπτούμε σειρές καταστάσεων (states), πράξεων / ενεργειών (actions) και ανταμοιβών (rewards) από πραγματική ή προσομοιωμένη αλληλεπίδραση με το περιβάλλον. Η μάθηση από αληθινή εμπειρία είναι εντυπωσιακή, γιατί δεν απαιτεί προηγούμενη γνώση των δυναμικών του περιβάλλοντος, ενώ μπορούμε επίσης να πετύχουμε ιδανική συμπεριφορά. Η μάθηση από προσομοιωμένη εμπειρία είναι επίσης πολύ ισχυρή. Παρόλο που και πάλι χρειάζεται ένα μοντέλο, αυτό πρέπει μόνο να παράγει τη δειγματοληψία των μεταβάσεων και όχι τις πλήρεις κατα-

νομές πιθανοτήτων όλων των δυνατών μεταβάσεων, που απαιτούνται από τις μεθόδους δυναμικού προγραμματισμού.

Οι μέθοδοι Monte Carlo αποτελούν έναν τρόπο λύσης του προβλήματος της επιβραβευμένης μάθησης, που βασίζεται στο μέσο όρο των αποτελεσμάτων (returns) δειγματοληψίας. Για να βεβαιωθούμε για την εγγυρότητα των returns, ορίζουμε τις μεθόδους Monte Carlo μόνο για εργασίες επεισοδίων. Δηλαδή, υποθέτουμε πως η εμπειρία είναι διαιρεμένη σε επεισόδια, τα οποία τελικά τερματίζουν αδιαφορώντας για το ποιες πράξεις έχουν επιλεγεί. Μόνο στην συμπλήρωση κάθε επεισοδίου εκτιμάται η αξία / τιμή (value) και αλλάζουν οι policies. Μπορεί άρα να θεωρηθεί πως οι μέθοδοι Monte Carlo είναι αυξητικές υπό την έννοια του episode-by-episode, αλλά όχι υπό την έννοια του step-by-step.

11.1 Αξιολόγηση της τακτικής κατά Monte Carlo (Monte Carlo policy evaluation)

Ξεκινούμε θεωρώντας τις μεθόδους Monte Carlo για τη μάθηση της συνάρτησης κατάστασης - αξίας για δεδομένη policy. Ένας προφανής τρόπος για να εκτιμηθεί η αξία της κατάστασης από την εμπειρία είναι να υπολογιστεί ο μέσος όρος των returns που παρατηρούνται μετά από επισκέψεις στην κατάσταση. Όσο περισσότερα returns σημειώνονται τόσο ο μέσος όρος θα πρέπει να συγκλίνει στην αναμενόμενη αξία. Αυτή η ιδέα υπογραμμίζει όλες τις μεθόδους Monte Carlo.

Συγκεκριμένα, έστω ότι θέλουμε να εκτιμήσουμε την $V^\pi(s)$, την τιμή μιας κατάστασης s για την policy π , με δεδομένο ένα σύνολο από επεισόδια λαμβανόμενα από συνεχή π , που περνούν μέσω της s . Κάθε εμφάνιση της κατάστασης s σε ένα επεισόδιο ονομάζεται “ μια επίσκεψη στην s ”. Η every-visit MC μέθοδος εκτιμά την $V^\pi(s)$, σαν το μέσο όρο των returns που ακολουθούν όλες τις επισκέψεις στην s σε ένα σύνολο επεισοδίων. Σε ένα δεδομένο επεισόδιο, η πρώτη φορά που κάποιος επισκέπτεται την s ονομάζεται “ πρώτη επίσκεψη στην s ”. Η first-visit MC μέθοδος υπολογίζει το μέσο όρο των returns που ακολουθούν τις πρώτες επισκέψεις στην s .

11.2 Εκτίμηση των αξιών των πράξεων κατά Monte Carlo (Monte Carlo estimation of action values)

Αν δεν είναι διαθέσιμο ένα μοντέλο, τότε είναι ιδιαίτερα χρήσιμο να εκτιμηθούν οι action values αντί των state values. Όταν υπάρχει μοντέλο, οι state values και μόνο είναι ικανές για να καθοριστεί η policy. Απλώς, η μία κοιτά μπροστά ένα βήμα και επιλέγει έτσι ποια λειτουργία οδηγεί στον καλύτερο συνδυασμό ανταμοιβής και επόμενης κατάστασης. Χωρίς μοντέλο, από την άλλη, οι state values μόνες τους δεν είναι επαρκείς. Μία πρέπει κατηγορηματικά να εκτιμήσει την value κάθε λειτουργίας, ώστε όλες οι values να μπορούν μετά να είναι σε θέση να προτείνουν μια policy. Επομένως, ένας από τους αρχικούς σκοπούς των μεθόδων Monte Carlo είναι η εκτίμηση της Q^* . Για να επιτευχθεί κάτι τέτοιο, θεωρούμε ένα άλλο policy evaluation πρόβλημα.

Το policy evaluation πρόβλημα για action values είναι η εκτίμηση της $Q^\pi(s, a)$, η εύρεση του αναμενόμενου return, όταν έχουμε ξεκινήσει από την κατάσταση s , κάνοντας χρήση της λειτουργίας a , οπότε φτάνουμε στην policy π . Οι Monte Carlo μέθοδοι και εδώ είναι οι ίδιες. Η every-visit MC μέθοδος υπολογίζει το ζευγάρι state-action ως το μέσο όρο των returns, που προέκυψαν μετά από επισκέψεις στην κατάσταση στην οποία επιλέχθηκε η λειτουργία. Η first-visit MC μέθοδος βρίσκει τον μέσο όρο των returns που πέρασαν την πρώτη φορά από κάθε επεισόδιο, όπου έγινε επίσκεψη στην κατάσταση και επιλέχθηκε η αντίστοιχη action.

Τέλος, αναφέρουμε πώς η εκτίμηση κατά Monte Carlo μπορεί να χρησιμοποιηθεί στον έλεγχο, δηλαδή στην προσέγγιση ιδανικών policies. Η όλη ιδέα σχετίζεται με την ιδέα της Γενικευμένης Επανάληψης Τακτικών (GPI). Σύμφωνα με την GPI διατηρείται και μια προσεγγιστική policy και μια προσεγγιστική value function. Η value function συνεχώς τροποποιείται για την καλύτερη προσέγγισή της για την τρέχουσα policy, και η policy συνεχώς βελτιώνεται σε σχέση με την τρέχουσα value function.

12 Μάθηση με χρονική διαφορά (temporal difference learning)

Αν κάποιος έπρεπε να προσδιορίσει μια ιδέα ως βασική και νέα στον τομέα της επιβραβευμένης μάθησης, αναμφίβολα αυτή θα ήταν η μάθηση με χρονική διαφορά (*Temporal Difference (TD) μάθηση*). Η TD μάθηση είναι ένας

συνδυασμός των ιδεών κατά Monte Carlo και των ιδεών του Δυναμικού Προγραμματισμού (DP). Οι TD μέθοδοι μπορούν να μάθουν κατευθείαν από μικρή εμπειρία, χωρίς κάποιο μοντέλο των δυναμικών του περιβάλλοντος. Όπως και στον Δυναμικό Προγραμματισμό, οι TD μέθοδοι ανανεώνουν τις εκτιμήσεις τους βασισμένες σε άλλες εκτιμήσεις μάθησης, χωρίς να περιμένουν για το τελικό αποτέλεσμα. Η σχέση μεταξύ των μεθόδων TD, DP και Monte Carlo είναι ένα επαναλαμβανόμενο θέμα στη θεωρία της επιβραβευμένης μάθησης.

12.1 Η TD Πρόβλεψη

Τόσο η TD μέθοδος όσο και η μέθοδος Monte Carlo χρησιμοποιούν την εμπειρία για να λύσουν το πρόβλημα της πρόβλεψης. Με δεδομένη κάποια εμπειρία από τη συγκεκριμένη policy π , και οι δύο μέθοδοι ανανεώνουν την εκτίμησή τους V από τη V^π . Αν γίνεται επίσκεψη σε μια μη τελική κατάσταση s_t σε χρόνο t , τότε και οι δύο μέθοδοι ανανεώνουν την εκτίμησή τους, $V(s_t)$, βασισμένες στο τι συμβαίνει μετά την επίσκεψη. Σε αντίθεση με την Monte Carlo, η οποία περιμένει μέχρι το τέλος του επεισοδίου, ώστε να σταματήσει η αύξηση σε $V(s_t)$ (μόνο τότε η R_t είναι γνωστή), οι TD μέθοδοι πρέπει να περιμένουν μόνο μέχρι το επόμενο χρονικό βήμα. Στον χρόνο $t + 1$ θέτουν αμέσως ένα στόχο και κάνουν μια χρήσιμη ανανέωση, χρησιμοποιώντας την παρατηρούμενη reward r_{t+1} και την εκτίμηση $V(s_{t+1})$. Η πιο απλή TD μέθοδος, γνωστή ως TD(0), είναι:

$$V(s_t) \leftarrow V(s_t) + a[r_{t+1} + gV(s_{t+1}) - V(s_t)]$$

Επειδή η TD μέθοδος βασίζει τις ανανεώσεις της κατά ένα μέρος σε μια υπάρχουσα εκτίμηση, χαρακτηρίζεται ως bootstrapping μέθοδος (μέθοδος αυτοδύναμης εκκίνησης), όπως ο δυναμικός προγραμματισμός.

Τελικά, οι TD μέθοδοι συνδυάζουν τη δειγματοληψία των Monte Carlo και το bootstrapping των DP.

12.2 Πλεονεκτήματα των TD μεθόδων πρόβλεψης

Οι TD μέθοδοι μαθαίνουν τις εκτιμήσεις τους βασισμένες κατά ένα μέρος σε άλλες εκτιμήσεις. Μαθαίνουν να μαντεύουν από εικασίες (bootstrapping). Αυτό τους προσφέρει πλεονεκτήματα σε σύγκριση με τις άλλες μεθόδους.

Καταρχήν οι TD μέθοδοι πλεονεκτούν των DP μεθόδων, γιατί δεν απαιτούν ένα μοντέλο του περιβάλλοντος.

Το επόμενο εμφανές πλεονέκτημά τους έναντι των μεθόδων Monte Carlo είναι πως πραγματοποιούνται με έναν τρόπο συνεχώς αυξητικό. Δηλαδή, ενώ στην περίπτωση των δεύτερων πρέπει να περιμένουμε μέχρι το τέλος του επεισοδίου, γιατί μόνο τότε το αποτέλεσμα είναι γνωστό, στην περίπτωση των πρώτων αρκεί να περιμένουμε μέχρι το επόμενο χρονικό βήμα.

Ευτυχώς οι TD μέθοδοι είναι και ορθές με την εξής έννοια: Εκτός του ότι είναι βολικό να μαθαίνουμε μια υπόθεση βασιζόμενοι σε μια άλλη, χωρίς να περιμένουμε το αποτέλεσμα, μπορούμε και να εγγυηθούμε σύγκλιση στη σωστή απάντηση.

Αν τόσο οι TD μέθοδοι όσο και οι MC μέθοδοι συγκλίνουν ασυμπτωτικά στις σωστές προβλέψεις, τότε αναρωτιέται κανείς ποια είναι η ταχύτερη, δηλαδή ποια μαθαίνει πιο γρήγορα και ποια κάνει πιο αποτελεσματική χρήση των περιορισμένων δεδομένων. Αυτή τη στιγμή δεν έχει αποδειχτεί μαθηματικά αν κάποια συγκλίνει ταχύτερα από την άλλη. Στην πράξη πάντως, οι TD μέθοδοι βρέθηκε ότι συγκλίνουν ταχύτερα από τις MC μεθόδους σε στοχαστικές εργασίες.

12.3 Η ιδανικότητα της TD(0)

Ας υποθέσουμε ότι είναι διαθέσιμο ένα πεπερασμένο σύνολο εμπειριών, έστω 10 επεισόδια ή 100 βήματα χρόνου. Στην περίπτωση αυτή, μια κοινή προσέγγιση με τις αυξητικές μεθόδους μάθησης είναι η επαναλαμβανόμενη παρουσίαση της εμπειρίας μέχρις ότου η μέθοδος να συγκλίνει σε μια απάντηση. Με δεδομένη μια προσεγγιστική value function, V , οι αυξήσεις υπολογίζονται για κάθε βήμα χρόνου t , στο οποίο έχουμε επίσκεψη σε μια μη τελική κατάσταση, όμως η value function αλλάζει μόνο μια φορά στο σύνολο όλων των αυξήσεων. Στη συνέχεια όλη η διαθέσιμη εμπειρία προωθείται ξανά με την καινούργια value function για να παραχθεί μια νέα γενική αύξηση και συνεχίζει μέχρι η νέα value function να συγκλίνει. Αυτό ονομάζεται μαζί ανανέωση (batch updating), γιατί οι ανανεώσεις γίνονται μόνο μετά την προώθηση κάθε πλήρους ομάδας (batch) εκπαιδευόμενης πληροφορίας.

Σύμφωνα με το batch updating, η TD(0) συγκλίνει ντετερμινιστικά σε μια μοναδική απάντηση ανεξάρτητα από την παράμετρο μέγεθος βήματος, α ,

εφόσον το α έχει διαλεγθεί να είναι αρκετά μικρό. Η constant- α MC μέθοδος επίσης συγκλίνει ντετερμινιστικά κάτω από τις ίδιες συνθήκες, αλλά σε διαφορετική απάντηση. Κατανοώντας τις δύο αυτές απαντήσεις, μπορούμε να βοηθηθούμε στο να καταλάβουμε τη διαφορά ανάμεσα στις δύο αυτές μεθόδους. Με μια φυσιολογική ανανέωση οι μέθοδοι αυτές δεν κατευθύνονται στις αντίστοιχες batch απαντήσεις, αλλά κατά μια έννοια κάνουν ένα βήμα προς αυτήν την κατεύθυνση.

12.4 Sarsa: On-Policy TD Control

Μπορούμε να χρησιμοποιήσουμε τις TD μεθόδους πρόβλεψης στο πρόβλημα του ελέγχου. Ως συνήθως χρησιμοποιούμε το πρότυπο GPI, μόνο που αυτή τη φορά στο τμήμα του υπολογισμού ή της πρόβλεψης εφαρμόζουμε TD μεθόδους. Όπως και στην περίπτωση των μεθόδων Monte Carlo, αντιμετωπίζουμε την ανάγκη να εξισορροπήσουμε την εκμετάλλευση και την εξερεύνηση. Οι προσεγγίσεις χωρίζονται σε δύο κατηγορίες: την on-policy και την off-policy. Η sarsa είναι μια on-policy TD μέθοδος ελέγχου.

Το πρώτο βήμα είναι να μάθουμε μια action-value function και όχι μια state-value function. Πιο συγκεκριμένα, σε μια on-policy μέθοδο πρέπει να εκτιμήσουμε την $Q^\pi(s, a)$ για την συμπεριφορά της τρέχουσας policy π και για όλες τις καταστάσεις s και τις πράξεις a . Θεωρούμε πως κάθε επεισόδιο αποτελείται από μια εναλλαγή ζευγαριών καταστάσεων και καταστάσεων-πράξεων. Ακόμη θεωρούμε μεταβάσεις από ζευγάρια καταστάσεων-πράξεων σε ζευγάρια καταστάσεων-πράξεων.

Οι ανανεώσεις γίνονται μετά από κάθε μετάβαση από μια μη τελική κατάσταση s_t . Αν η s_{t+1} είναι τελική, τότε η $Q(s_{t+1}, a_{t+1})$ ορίζεται ως μηδενική. Αυτός ο κανόνας χρησιμοποιεί κάθε στοιχείο μιας πεντάδας γεγονότων, $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$, που πραγματοποιεί μια μετάβαση από ένα ζευγάρι κατάσταση-πράξης στο επόμενο. Αυτή η πεντάδα είναι που δίνει το όνομα sarsa στον αλγόριθμο.

12.5 Q-learning: Off-Policy TD Control

Μια από τις πιο σημαντικές ανακαλύψεις στην επιβραβευμένη μάθηση ήταν η ανάπτυξη ενός off-policy TD αλγορίθμου ελέγχου, γνωστού ως Q-learning. Η πιο απλή του μορφή, ο 1-step Q-learning ορίζεται ως εξής:

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

Στην περίπτωση αυτή, η εκπαιδευμένη action-value function, Q , προσεγγίζει απευθείας την Q^* , τη βέλτιστη action-value function, ανεξάρτητα από την policy που ακολουθεί. Κάτι τέτοιο απλοποιεί σημαντικά την ανάλυση του αλγορίθμου και ενισχύει τις σύντομες αποδείξεις σύγκλισης. Η policy εξακολουθεί βέβαια να επιδρά στο ότι καθορίζει σε ποια ζευγάρια καταστάσης-πράξης γίνεται επίσκεψη και ανανέωση. Ωστόσο ό,τι απαιτείται για μια σωστή σύγκλιση είναι πως όλα τα ζευγάρια συνεχίζουν να ανανεώνονται. Πάντως, αυτό είναι μια ελάχιστη απαίτηση, μιας και κάθε μέθοδος που εγγυάται πως θα βρει τη βέλτιστη συμπεριφορά στη γενική περίπτωση πρέπει να απαιτεί κάτι τέτοιο.

13 Ίχνη καταλληλότητας (Eligibility Traces)

Τα Ίχνη ή Σημάδια Καταλληλότητας (*Eligibility Traces*) είναι ένας από τους βασικούς μηχανισμούς της επιβραβευμένης μάθησης. Για παράδειγμα, στον δημοφιλή TD(λ) αλγόριθμο, ο λ αναφέρεται στη χρήση ενός eligibility trace. Σχεδόν οποιαδήποτε temporal-difference (TD) μέθοδος, για παράδειγμα η Q-learning ή η Sarsa, μπορεί να συνδυαστεί με τα eligibility traces ώστε να προκύψει μια γενική μέθοδος, η οποία θα μπορεί να μαθαίνει πιο αποτελεσματικά.

Υπάρχουν δύο τρόποι για να μελετήσουμε τα eligibility traces. Η μια άποψη, η πιο θεωρητική, υποστηρίζει ότι αποτελούν μια γέφυρα από τις TD στις Monte Carlo μεθόδους. Όταν οι TD μέθοδοι γεμίζουν με eligibility traces, παράγουν μια οικογένεια μεθόδων γεμίζοντας ένα φάσμα, που έχει από τη μια τις μεθόδους Monte Carlo και από την άλλη τις 1-step TD μεθόδους. Ενδιαμέσως βρίσκονται μέτριες μέθοδοι, που όμως συχνά είναι καλύτερες από οποιαδήποτε ακραία. Με την έννοια αυτή, τα eligibility traces ενώνουν τις μεθόδους TD και Monte Carlo κατά έναν πολύτιμο και αποκαλυπτικό τρόπο.

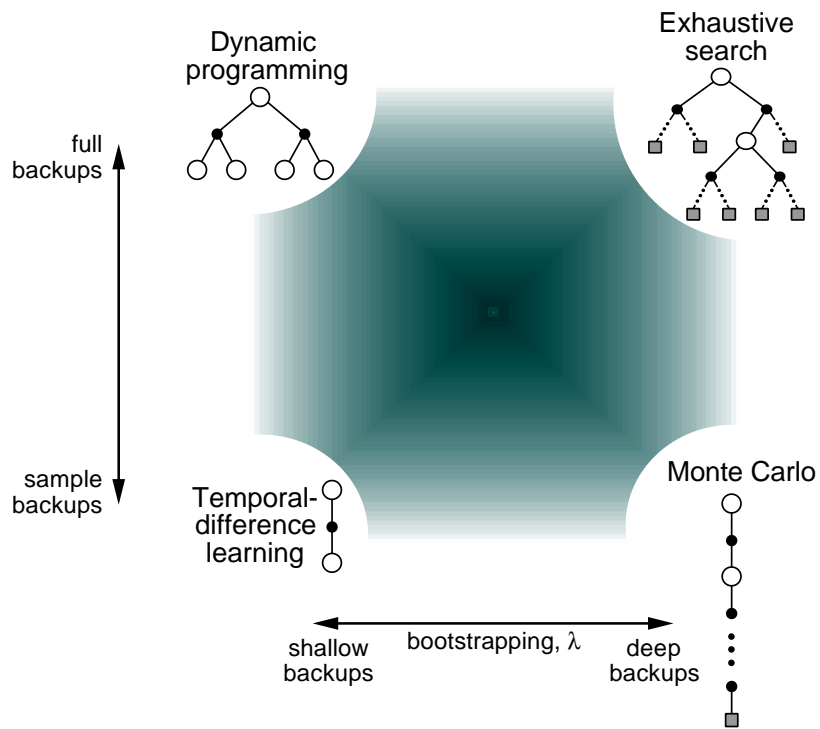
Η άλλη άποψη μελέτης των eligibility traces είναι πιο μηχανιστική. Από αυτήν την πλευρά, ένα eligibility trace είναι μια προσωρινή σημείωση ενός συμβάντος, όπως η επίσκεψη μιας κατάστασης ή η πραγματοποίηση μιας λειτουργίας. Το σημάδι ή ίχνος (trace) σημειώνει στη μνήμη παραμέτρους συνδεδεμένες με το γεγονός ως κατάλληλες να υποστούν αλλαγές εκμάθησης. Όταν συμβαίνει ένα TD λάθος, μόνο οι κατάλληλες καταστάσεις ή λειτουργίες εκχωρούνται ως υπεύθυνες για το λάθος. Έτσι, τα eligibility traces βοηθούν στη γεφύρωση του χάσματος μεταξύ γεγονότων και εκπαιδευόμενης πληροφορίας.

Η πιο θεωρητική άποψη των eligibility traces ονομάζεται forward view (θεώρηση προς τα εμπρός), και η πιο μηχανιστική άποψη backward view (θεώρηση προς τα πίσω). Η forward view είναι πιο χρήσιμη για να καταλάβουμε τι υπολογίζεται από τις μεθόδους που χρησιμοποιούν τα eligibility traces, ενώ η backward view είναι πιο κατάλληλη για την ανάπτυξη διαίσθησης στους ίδιους τους αλγόριθμους.

14 Ενοποιημένη θεώρηση των μεθόδων επιβραβευμένης μάθησης

Ο σωστότερος ίσως τρόπος να δει κανείς την επιβραβευμένη μάθηση είναι όχι ως μια συλλογή μεμονωμένων μεθόδων, αλλά ως ένα συνεκτικό σύνολο από ιδέες που διασταυρώνονται με συγκεκριμένες μεθόδους. Όλες οι βασικές μέθοδοι επιβραβευμένης μάθησης αξιοποιούν τις ακόλουθες ιδέες: Έχουν ως στόχο την εκτίμηση των συναρτήσεων αξίας, λειτουργούν αποθηκεύοντας τιμές ταυτόχρονα με την πραγματική ή τις πιθανές διαδοχές καταστάσεων και ακολουθούν τη γενική στρατηγική της generalized policy iteration - GPI, διατηρώντας μία προσεγγιστική συνάρτηση αξιών και μία προσεγγιστική τακτική και προσπαθώντας συνεχώς να βελτιώσουν τη μία βάσει της άλλης.

Δύο από τα πιο σημαντικά στοιχεία διαφοροποίησης των μεθόδων απεικονίζονται στο σχήμα 3. Η διαφοροποίηση έχει να κάνει με το είδος του backup που χρησιμοποιείται για τη βελτίωση της συνάρτησης αξίας και το οποίο μπορεί να είναι είτε sample backup (βασισμένο σε μία διαδοχή), είτε full backup (βασισμένο σε μία κατανομή πιθανών διαδοχών). Βέβαια μόνο στην περίπτωση των full backups είναι απαραίτητο ένα μοντέλο του περιβάλλοντος. Άλλο ένα σημείο διαφοροποίησης έχει να κάνει με το βάθος του backup, δηλαδή το βαθμό που η εκκίνηση είναι αυτοδύναμη. Επίσης σημείο διαφοροποίησης είναι και αυτό της προσέγγισης της συνάρτησης, η οποία μπορεί να γίνει με μεθόδους γραμμικές ή μη. Η διάσταση αυτή θα μπορούσε να απεικονισθεί ως κάθετη στο επίπεδο του παραπάνω σχήματος. Τέλος σημαντικό σημείο διαφοροποίησης μεταξύ των μεθόδων επιβραβευμένης μάθησης είναι και το αν πρόκειται για on-policy ή για off-policy μεθόδους. Στην πρώτη περίπτωση ο πράκτορας μαθαίνει τη συνάρτηση αξίας για την τακτική που ακολουθεί εκείνη τη στιγμή, ενώ στη δεύτερη περίπτωση για την τακτική που εκείνη τη στιγμή θεωρεί καλύτερη.



Σχήμα 3: Σύγκριση βασικών μεθόδων επιβραβευμένης μάθησης.

15 Συμπεράσματα

Η Επιβραβευμένη Μάθηση βασίζεται στην ιδέα της μάθησης από την αλληλεπίδραση με το περιβάλλον, από τις συνέπειες πράξεων, παρά από ρητή διδασκαλία. Οι μέθοδοι Επιβραβευμένης Μάθησης προορίζονται για την αντιμετώπιση προβλημάτων μάθησης και λήψης αποφάσεων σαν και αυτά που αντιμετωπίζουν άνθρωποι και ζώα στην καθημερινή τους ζωή. Ήδη υπάρχει μία ποικιλία τεχνικών επιβραβευμένης μάθησης που δουλεύει αποτελεσματικά για μια ποικιλία μικρών προβλημάτων. Ωστόσο λίγες από τις τεχνικές αυτές μπορούν να χρησιμοποιηθούν για μεγάλα, γενικευμένα, αυθαίρετα και πολύπλοκα προβλήματα. Τέτοια προβλήματα απαιτούν -προς το παρόν τουλάχιστον- την εισαγωγή μεροληψίας κατά τη διαδικασία της μάθησης, παρά τεχνικές μάθησης που ξεκινούν από το μηδέν (*tabula rasa*), προκειμένου να λυθούν αποδοτικά.

Αναφορές

- [1] Tsitsiklis J. N Bertsekas D. P. *Neural Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.
- [2] Thomas G. Dietterich. Hierarchical reinforcement learning. Department of Computer Science, Oregon State University, Corvallis, Oregon 97331, 1999.
- [3] Andrew W. Moore Leslie Pack Kaelbling, Michael L. Littman. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996.
- [4] Stephanie S. Harmon Mance E. Harmon. Reinforcement learning: A tutorial, 1996.
- [5] A. Barto R. Crites. Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 1998.
- [6] Andrew G. Barto Richard S. Sutton. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
- [7] T.K. Das A. Gosavi S. Mahadevan, N. Marchalleck. Self-improving factory simulation using continuous-time average-reward reinforcement learning. Στο *Proceedings of the 14th International Conference on Machine Learning*, 1996.
- [8] David Cohn Satinder Singh, Peter Norvig. How to make software agents do the right thing: An introduction to reinforcement learning. Adaptive Systems Group, Harlequin Inc., 1996.
- [9] Rich Sutton. Reinforcement learning faq: Frequently asked questions about reinforcement learning, 2001.
- [10] Ροβέρτος - Ε. Κινγκ. *Υπολογιστική Νοημοσύνη στον Έλεγχο Συστημάτων*. Εκδόσεις Π. Τραυλός - Ε. Κωσταράκη, Αθήνα, 1998.