ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Δημιουργία συστάσεων σε διάλογο με χρήση contextual bandits

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Λεωνίδα Αβδελά

Επιβλέπων: -Εισάγετε το όνομα, αρχικό πατρώνυμο και επίθετο του επιβλέποντα-
-Εισάγετε τον τίτλο του επιβλέποντα-

Αθήνα, Ιανουάριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Δημιουργία συστάσεων σε διάλογο με χρήση contextual bandits

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Λεωνίδα Αβδελά

Επιβλέπων: -Εισάγετε το όνομα, αρχικό πατρώνυμο και επίθετο του επιβλέποντα-
-Εισάγετε τον τίτλο του επιβλέποντα-

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την -εισάγετε ημερομηνία-.

.....	.....	.....
-Εσαγετε Ονοματεπωνυμο-	-Εσαγετε Ονοματεπωνυμο-	-Εσαγετε Ονοματεπωνυμο-
-Εσαγετε τίτλο-	-Εσαγετε τίτλο-	-Εσαγετε τίτλο-

Αθήνα, Ιανουάριος 2023.

.....

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

Περίληψη

Η δημιουργία συστάσεων είναι ένα πρόβλημα στο οποίο έχουν δείξει ενδιαφέρον τόσο η ακαδημαική κοινότητα, όσο και η βιομηχανία. Επιτυχημένες εφαρμογές έχουν χτιστεί πάνω στην δημιουργία καλών συστάσεων, όπως ο αλγόριθμος του Netflix. Η άνθιση προέρχερται κυρίως λόγο του μεγάλου όγκου πληροφοριών που υπάρχουν στο διαδίκτυο, του οποίου η αναζήτηση και επιμέλεια από μεμονομένα άτομα είναι πρακτικά αδύνατη.

Επιπλέον, η ενισχυτική μάθηση είναι ένας απο τους πλέον διαδεδομένους τρόπους εκπαίδευσης πρακτόρων, κυρίως σε περιβάλλοντα που το τελικό αποτέλεσμα γίνεται γνωστό μετά από πολλά βήματα, και δεν υπάρχει γνωστή βέλτιστη λύση για κάθε βήμα.

Τέλος η χρήση μηχανική μάθησης για την παραγωγή και την κατανόηση κειμένου είναι ένας κλάδος ο οποίος έχει δει μεγάλη άνθιση τα τελευταία λίγα χρόνια

Η τρέχουσα διπλωματική ασχολήται με την δημιουργία ενός συστήματος συστάσεων, το οποίο δουλεύει παράλληλα με ένα διαλογικό σύστημα, το οποίο προτείνει θέματα συζήτησης στον χρήστη. Η επιλογή των πρακτόρων έγινε με βάση την γνώστη ότι το περιβάλλον εργασίας περιείχε μειωμένα δεδομένα, οπότε κρίθηκε η χρήση τεχνικών ενισχυτικής μάθησης ως η βέλτιστη λύση.

Λέξεις Κλειδιά

TODO

Abstract

TODO

Keywords

TODO

Ευχαριστίες

Ευχαριστώ την οικογένεια μου και τους καθηγητές μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

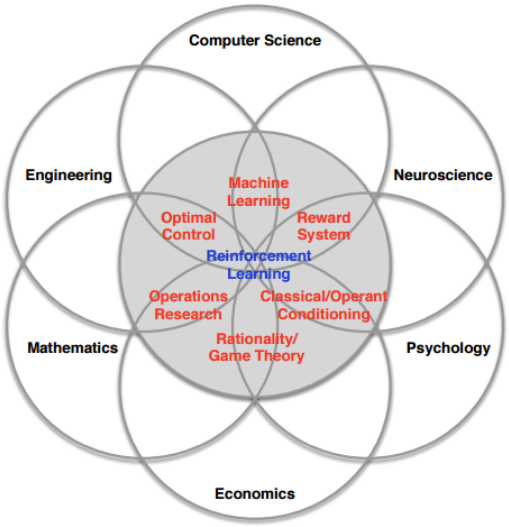
Ενισχυτική Μάθηση

1.1 Γενικά

Η ενισχυτική μάθηση (ΕΜ) (Reinforcement Learning (RL)) είναι είναι ένας γενικός όρος που έχει δοθεί σε μια οικογένεια τεχνικών στις οποίες ένα σύστημα προσπαθεί να μάθει μέσα από την άμεση αλληλεπίδραση με το περιβάλλον [2]. Είναι τομέας της τεχνητής νοημοσύνης και, πιο συγκεκριμένα, της μηχανικής μάθησης.

Πιο συγκεκριμένα, η ΕΜ είναι η διαδικασία κατα την οποία ένας πράκτορας (agent) αλληλεπιδρά με το περιβάλλον του, και μαθαίνει τι να κάνει, παρατηρώντας τις συνέπειες των πράξεων του. Ο πράκτορας δεν δίνεται πληροφορίες σχετικά με το ποίες ενέργειες (actions) να επιλέξει, αλλά πρέπει να ανακαλύψει ποιες ενέργειες προσφέρουν την μέγιστη ανταμοιβή (reward), δοκιμάζοντας τες [11]. Επιπλέον, σε πολλές περιπτώσεις οι ενέργειες του πράκτορα δεν επηρεάζουν μόνο την άμεση ανταμοιβή που θα πάρει, αλλά και από την ανταμοιβή στην επόμενη κατάσταση, και πιθανώς και όλες τις επόμενες ανταμοιβές. Ετσι,μπορεί να υπάρξουν καταστάσεις που ο πρακτορας θα πρέπει να θυσιάσει την άμεση ανταμοιβή για να αποκτήσει καλύτερες ανταμοιβές μακροπρόθεσμα. Σύμφωνα με τα παραπάνω, η ΕΜ στοχεύει να λύσει προβλήματα μέσω τεχνικών *δοκιμής-και-λάθους (trial-and-error)* σε περιβάλλοντα με *καθυστερημένες ανταμοιβές*.

Μια κομβική ιδέα, πάνω στην οποία στηρίζεται η ΕΜ, είναι η υπόθεση της ανταμοιβής (reward hypothesis), η ιδέα ότι κάθε στόχος μπορεί να εκφραστεί ως η μεγιστοποίηση της αναμενόμενης αξίας του σωρευτικού (cummulative) αθροίσματος ενός μονοδιάστατου σήματος. Με απλά λόγια, η υπόθεση θέτει την ιδέα ότι κάθε στόχος μπορεί να εκφραστεί σαν την μεγιστοποίηση μιας ανταμοιβής. Η ανταμοιβή αυτή δεν χρειάζεται απαραίτητα να είναι θετικός αριθμός, αλλά ακόμα και για αρνητικές τιμές της, συνεχίζουμε να καλούμε τον όρο ανταμοιβή. Για παράδειγμα, αν ο στόχος είναι η έξοδος από ένα λαβύρινθο, η ανταμοιβή μπορεί να είναι αρνητική σε κάθε βήμα μέχρι την έξοδο, οπού και γίνεται 0. Τότε ο στόχος τελικά είναι η ελαχιστοποίηση της απόλυτης τιμής της ανταμοιβής, δηλαδή η έξοδος στα λιγότερα βήματα.



Σχήμα 1.1: Τα πρόσωπα της ενισχυτικής μάθησης

Το πεδίο της ΕΜ έχει τις ρίζες του σε δύο περιοχές. Η πρώτη είναι η συμπεριφορική ψυχολογία, από όπου προέρχεται το παράδειγμα της δοκιμής-και-λάθους, και η δεύτερη είναι η περιοχή του βέλτιστου ελέγχου, από όπου η ΕΜ δανείζεται τον μαθηματικό φορμαλισμό (κυρίως τον δυναμικό προγραμματισμό) που υποστηρίζει το πεδίο. Είναι σημαντικό να γνωρίζουμε ότι η ΕΜ βρίσκεται στην τομή πολλών διαφορετικών επιστημονικών πεδίων, οι οποίοι φαίνονται στο Σχήμα 1.1[5]. Όλα αυτά τα πεδία προσεγγίζουν ένα παρόμοιο πρόβλημα, αλλά απο διαφορετική σκοπιά και με διαφορετικές παραμέτρους.

Η ΕΜ πολλές φορές συγχέεται με τις άλλες τεχνικές μηχανικής μάθησης, την επιβλεπόμενη και την μη επιβλεπόμενη μάθηση, παρόλο που έχει αρκετά σημαντικές διαφορές.

Αρχικά, η κύρια διαφορά μεταξύ της επιβλεπόμενης μάθησης (supervised learning) και της ΕΜ είναι ότι στην επιβλεπόμενη μάθηση, το μοντέλο εκπαιδεύεται πάνω σε δείγματα (samples) και ετικέτες (labels), και κάθε πρόβλεψη θεωρείται μοναδικό γεγονός. Στόχος είναι η προσέγγιση μιας άγνωστης συνάρτησης με βάση τα δεδομένα. Αντίθετα, στην ΕΜ, μπορούν να υπάρχουν πολλά βήματα πριν ο πράκτορας μάθει αν η απόφαση που πήρε ήταν σωστή, και είναι πιθανό να μην μάθει ποτέ ποια ήταν η αληθής/βέλτιστη τιμή. Το μόνο που παρατηρεί είναι η επίδραση που είχαν οι πράξεις του στο περιβάλλον.

Όσον αφορά την μη επιβλεπόμενη μάθηση, μπορεί αρχικά να φαίνεται παρόμοια με την ΕΜ. Όμως, στόχος της μη επιβλεπόμενης μάθησης είναι η εύρεση της κρυμμένης δομής δεδομένων τα οποία δεν έχουν ετικέτες. Η ΕΜ έχει διαφορετικό στόχο, ο οποίος είναι η μεγιστοποίηση του σήματος ανταμοιβής. Παρόλο που η εύρεση δομής είναι πολύ σημαντική και στην ΕΜ ώστε να μπορέσει ο πράκτορας να επιλέξει τις κατάλληλες κινήσεις, αυτό από μόνο του δεν

επιτυγχάνει τον στόχο της ΕΜ.

Ενα από τα κύρια προβλήματα της ΕΜ, που δεν συναντάται στις άλλες μορφές μηχανικής μάθησης είναι ο συμβιβασμός μεταξύ εξερεύνησης και εκμετάλλευσης. Για να αποκτήσει ο πράκτορας μεγάλη ανταμοιβή θα προτιμήσει τις ενέργειες που δοκίμασε στο παρελθόν και του προσέφεραν μεγαλύτερη ανταμοιβή. Αλλά για να ανακαλύψει τέτοιες πράξεις, πρέπει να δοκιμάσει πράξεις που δεν έχει δοκιμάσει στο παρελθόν. Έτσι ο πράκτορας πρέπει να *εκμεταλλευτεί* το τι έχει ήδη βιώσει ώστε να αποκτήσει ανταμοιβές, αλλά πρέπει και να *εξερευνήσει*, ώστε να πάρει καλύτερες αποφάσεις στο μέλλον. Έτσι το δίλημμα είναι ποια από τις δύο στρατηγικές να επιλέξει κάθε φορά, καθώς καμία δεν μπορεί επιδιωχθεί αποκλειστικά, έτσι ώστε να επιτευχθεί ο στόχος του πράκτορα. Ως αποτέλεσμα, ο πράκτορας πρέπει να δοκιμάσει διάφορες κινήσεις και προοδευτικά να προτιμήσει αυτές που θεωρεί καλύτερες.Καθώς πολλές από τα προβλήματα που θέλουμε να λύσουμε με ΕΜ είναι στοχαστικά, ο πράκτορας πρέπει να δοκιμάσει κάθε πράξη πολλές φορές για να πάρει μια αξιόπιστη εκτίμηση της προσδοκώμενης ανταμοιβής.

1.2 Στοιχεία της ΕΜ

Πέρα από την πράκτορα και το περιβάλλον, υπάρχουν ακόμα τέσσερα κύρια στοιχεία σε ένα σύστημα ΕΜ. Αυτά είναι:

- Η πολιτική (policy), η οποία ορίζει την συμπεριφορά του πράκτορα για μια δοθείσα χρονική στιγμη. Με απλά λόγια, η πολιτική είναι μια χαρτογράφηση από τις καταστάσεις που αντιλαμβάνεται ο πράκτορας στις ενέργειες που παίρνει σε αυτές τις καταστάσεις (πιο συγκεκριμένα στις πιθανότητες αυτών των ενεργειών). Οι πολιτικές μπορεί να είναι και στοχαστικές, προσδιορίζοντας μια πιθανότητα για κάθε ενέργεια.

- Το σήμα ανταμοιβής, το οποίο αναφέρθηκε και νωρίτερα. Το σήμα αυτό προσδιορίζει στόχος ενός προβλήματος ΕΜ. Σε κάθε χρονικό βήμα, το περιβάλλον στέλνει στον πράκτορα έναν αριθμό, την ανταμοιβή. Ο μόνος στόχος του πράκτορα είναι να μεγιστοποιήσει την συνολική ανταμοιβή που λαμβάνει μακροπρόθεσμα. Έτσι το σήμα της ανταμοιβής προσδιορίζει ποιά είναι τα καλά και τα κακά γεγονότα για τον πράκτορα. Είναι σημαντικό η ανταμοιβή να προσδιορίζει ακριβώς το τι θέλουμε να πετύχουμε. Η ανταμοιβή δεν πρέπει να περιέχει πληροφορίες για το πώς θα πετύχουμε τον στόχο. Αυτές οι πληροφορίες μπορούν να τοποθετηθούν μέσα στην πολιτική ή στην συνάρτηση αξίας. Το σήμα είναι η κύρια βάση λόγω της οποίας αλλάζει η πολιτική. Αν ο πράκτορας επιλέξει μια ενέργεια με μικρή ανταμοιβή, τότε η πολιτική του ίσως να αλλάξει για να επιλέξει κάποια άλλη ενέργεια, όταν υπάρξει η ίδια κατάσταση στο μέλλον. Γενικά, οι ανταμοιβές μπορεί να είναι στοχαστικές συναρτήσεις της κατάστασης του περιβάλλοντος και των ενεργειών που επιλέχθηκαν.

- Η συνάρτηση αξίας κάθε κατάστασης (value function). Αντίθετα από το σήμα ανταμοιβής που μας επιστρέφει το τί είναι καλό άμεσα, η συνάρτηση αξίας προσδιορίζει τι

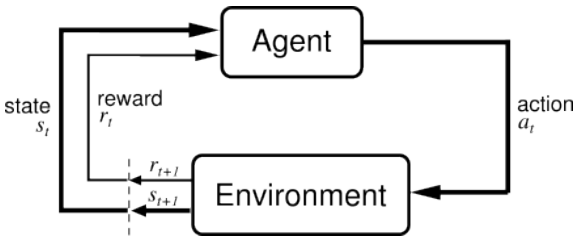
είναι καλό μακροπρόθεσμα. Σε γενικές γραμμές, η αξία μιας κατάστασης είναι η συνολική ανταμοιβή που μπορεί να περιμένει να αποκτήσει ένας πράκτορας στο μέλλον, ξεκινώντας από την συγκεκριμένη κατάσταση. Έτσι, ενώ οι ανταμοιβές προσδιορίζουν την άμεση και εσωτερική επιθυμητότητα των καταστάσεων του περιβάλλοντος, η αξία υποδεικνύει την μακροπρόθεσμη επιθυμητότητα των καταστάσεων, παίρνοντας υπόψιν τις καταστάσεις που θα ακολουθήσουν και τις διαθέσιμες ανταμοιβές σε αυτές τις καταστάσεις. Έτσι, μια κατάσταση με χαμηλή ανταμοιβή, μπορεί να έχει μεγάλη αξία γιατί οδηγεί σε καταστάσεις με μεγαλύτερες ανταμοιβές. Ετσι η συνάρτηση αξίας προσδιορίζει πόσο καλό είναι για ένα πράκτορα να είναι στην συγκεκριμένη κατάσταση.

- Προαιρετικά,ένα μοντέλο του περιβάλλοντος (model). Το μοντέλο ενός συστήματος ΕΜ μιμήται την συμπεριφορά του συστήματος, ή πιο γενικά, επιτρέπει την δημιουργία συμπερασμάτων για το πώς θα συμπεριφερθεί το περιβάλλον. Τα μοντέλα χρησιμοποιούνται για σχεδιασμό (planning), δηλαδή την επίλογή της σειράς των δράσεων παίρνοντας υπόψιν πιθανές μελλοντικες καταστάσεις, πριν τις βιώσει ο πράκτορας. Μέθοδοι επίλυσης προβλημάτων ΕΜ που χρησιμοποιούν μοντέλα και σχεδιασμό λέγονται μέθοδοι βασισμένοι σε μοντέλα (model-based). Αν οι μέθοδοι δεν έχουν μοντέλο, δηλαδή μέθοδοι που μαθαίνουν ρητά μέσω δοκιμής-και-λάθους, λέγονται μέθοδοι χωρίς μοντέλο (model-free).

Πιο φορμαλιστικά, σε ένα περιβάλλον ΕΜ, ένας αυτόνομος πράκτορας, ελεγχόμενος από ένα αλγόριθμο μηχανικής μάθησης, παρατηρεί μια κατάσταση s_t από το περιβάλλον του σε ένα χρονικό βήμα t. Οι χρονικές στιγμές στην περιγραφή αυτή είναι διακριτές, δηλαδή $\tau = 0, 1, 2, \ldots$, αλλά θα μπορούσαν να είναι και συνεχείς, χώρις μεγάλες διαφορές. Οι καταστάσεις προέρχονται από τον χώρο καταστάσεων $\mathcal{S}$. Ο πράκτορας αλληλεπιδρά με το περιβάλλον επιλέγοντας μια ενέργεια a_t με βάση την κατάσταση s_t, επιλεγμένη από ένα χώρο ενεργειών $\mathcal{A}(s)$. Όταν ο πράκτορας εκτελέσει την ενέργεια, τότε τόσο το περιβάλλον, μεταβαίνει σε μια νέα κατάσταση s_{t+1}, με βάση την τρέχουσα κατάσταση και την επιλεγμένη ενέργεια [8]. Σε κάθε τριπλέτα (κατάστασης, ενέργειας, νέας κατάστασης), αντιστοιχεί μια πιθανότητα μετάβασης $\Pr\{S_{t+1} = s' | S_t = s, A_t = a\}$. Σε κάθε κατάσταση,ο πράκτορας μπορεί είναι να παρατηρήσει την πλήρη δυναμική του περιβάλλοντος ή μέρος της. Ο πράκτορας επίσης λαμβάνει και μια μονοδιάστατη ανταμοιβή R_t η οποία προέρχεται από την τριπλέτα (κατάστασης, ενέργειας, νέας κατάστασης), και συμβολίζεται ως $R(s_{t-1}, a_{t-1}, s_t)$. Αυτή η ανταμοιβή δεν είναι γνωστή στον πράκτορα στην αρχή και δρα ως μια μορφή ανατροφοδότησης για τις δράσεις του πράκτορα. Αυτή η διαδικασία αναπαριστάται οπτικά στο Σχήμα 1.2.

Συνήθως ο πράκτορας διατηρεί μια εσωτερική κατάσταση, η οποία περιλαμβάνει κομμάτια της κατάστασης του περιβάλλοντος τα οποία θεωρούνται σημαντικά, καθώς και άλλες πληροφορίες. Σε αυτή την εσωτερική κατάσταση, ο πράκτορας διατηρεί μια αντιστοίχηση μεταξύ κατάστασης και ενέργειας, η οποία συμβολίζεται ως $\Pr(a_t | s_t)$.

Η βέλτιστη σειρά ενεργειών προσδιορίζεται από τις ανταμοιβές που προμηθεύει το περιβάλλον. Ο τελικός στόχος του πράκτορα είναι να μάθει μια πολιτική π, η οποία μεγιστοποιεί την αναμενόμενη απόδοση (σωρευτική, εκπτωθείσα (discounted) ανταμοιβή). Δοθείσας μιας



Σχήμα 1.2: Αλληλεπίδραση πράκτορα και περιβάλλοντος

κατάστασης η πολιτική αποφασίζει την επόμενη ενέργεια την οποία θα κάνει ο πράκτορας. Μια βέλτιστη πολιτική είναι η πολιτική η οποία μεγιστοποιεί την αναμενόμενη απόδοση στο συγκεκριμένο περιβάλλον.

Πέρα απο την περιγραφή του συστήματος ΕΜ με βάση την ύπαρξη ή όχι μοντέλου του περιβάλλοντος, ένας άλλος τρόπος να περιγράψουμε τα συστήματα ΕΜ είναι με βάση το περιβάλλον στο οποίο βρίσκονται οι πράκτορες. Η μία περίπτωση είναι να είναι αυτό το περιβάλλον πλήρως παρατηρήσιμο, δηλαδή ο πράκτορας μπορεί να παρατηρήσει κάθε πληροφορία για την δυναμική του περιβάλλοντος. Αυτό φυσικά δεν σημαίνει ότι κάθε παρατήρηση θα είναι χρήσιμη. Έτσι η κατάσταση του πράκτορα θα είναι το υποσύνολο των παρατηρήσεων που είναι χρήσιμες. Αυτά τα περιβάλλοντα ικανοποιούν την Μαρκοβιανή ιδιότητα. Δηλαδή, για κάθε κατάσταση, το μέλλον εξαρτάται μόνο από την τρέχουσα κατάσταση και όχι τις προηγούμενες. Όταν ισχύει αυτή η ιδιότητα τότε μπορούμε να μοντελοποιήσουμε το πρόβλημα ως μια Μαρκοβιανή Διαδικασία Αποφάσεων (Markov Decision Process). Αντίθετα, υπάρχουν περιβάλλοντα που δεν είναι πλήρως παρατηρήσιμα, όπως για παράδειγμα ένα δωμάτιο μέσα στο οποίο κινειται ένα ρομπότ. Σε αυτή την περίπτωση, το ρομπότ δεν γίνεται σε κάθε κίνηση του να γνωρίζει τα πάντα για το περιβάλλον γιατί υπάρχουν πάρα πολλές παράμετροι.

Η Μαρκοβιανή ιδιότητα ορίζεται ως:

$$p(r, s|S_t, A_t) = p(r, s|\mathcal{H}_t, A_t) \tag{1.1}$$

το οποίο σημαίνει ότι η πιθανότητα να βρεθούμε στην κατάσταση s με ανταμοιβή r, αν γνωρίζουμε ολόκληρη την ιστορία της αλληλεπίδρασης του πράκτορα με το περιβάλλον και παίρνοντας και κάνοντας την ενέργεια A_t (αριστερό μέρος) είναι ίδια με την πιθανότητα να βρεθούμε στην κατάσταση s με ανταμοιβή r γνωρίζοντας μόνο την τελευταία κατάσταση S_t στην οποία βρισκόταν ο πράκτορας και την ενέργεια A_t που έκανε (δεξί μέρος).

Σε μια Μαρκοβιανή Διαδικασία Αποφάσεων δημιουργούμε μια εκτίμηση της βέλτιστης $q_*(s, \alpha)$ κάθε ενέργειας α σε κάθε κατάσταση s ή μια εκτίμηση της αξίας $u_*(s)$ κάθε κατάστασης δεδομένης μιας βέλτιστης επιλογής ενεργειών.

1.3 Μαρκοβιανές Διαδικασίες Αποφάσεων

Για την καλύτερη κατανόηση της ΕΜ, είναι χρήσιμο να περιορίσουμε το πρόβλημα στην μορφή του που είναι μια ΜΔΑ. Συγκεκριμένα, σε διακριτό χρόνο, η αρχή της πορείας ενός πράκτορα θα είναι

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2$$

Σε μια πεπερασμένη ΜΔΑ, η κατάστασεις, οι ενέργειες και οι ανταμοιβές είναι τα σύνολα $(\mathcal{S}, \mathcal{A}, \mathcal{R})$ με πεπερασμένο αριθμό στοιχείων. Σε αυτή την περίπτωση οι τυχαίες μεταβλητές R_t και S_t έχουν μια καλά ορισμένη διακριτή πιθανότητα, η οποία εξαρτάται από την προηγούμενη κατάσταση και ενέργεια. Έτσι η δυναμική της ΜΔΑ ορίζεται από την συνάρτηση, η οποία λόγω της Μαρκοβιανής ιδιότητας προσδιορίζει πλήρως την δυναμική του συστήματος.

$$p(s', r|s, a) = \Pr S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a \tag{1.2}$$

Όσον αφορά το σήμα ανταμοιβής, ο στόχος είναι η μεγιστοποίηση του σε βάθος χρόνου. Αυτό αναφέρεται στην βιβλιογραφία ως απόδοση (return) οπως αναφέρθηκε και νωρίτερα, η οποία πολύ συχνά αναπαριστάται ως G_t.

Ανάλογα με το αν τερματίζει ένα πρόβλημα, αυτό μπορεί να θεωρηθεί επεισοδικό ή συνεχές. Σε ένα επεισοδικό πρόβλημα, υπάρχει πάντα μια τελική κατάσταση (π.χ. η έξοδος ενός λαβυρίνθου). Σε ένα συνεχές πρόβλημα, δεν υπάρχει αυτός ο διαχωρισμός σε επεισόδια, αλλά η αλληλεπίδραση συνεχίζεται ατέρμονα. Για να μπορέσουμε να υπολογίσουμε την απόδοση ακόμα και σε συνεχή προβλήματα, συνήθως την ορίζουμε ως

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1.3}$$

όπου $0 \leq \gamma \leq 1$ είναι μια παράμετρος η οποία ονομάζεται παράγοντας έκπτωσης. Ο παράγοντας αυτός επηρεάζει το πόσο σημασία έχουν οι μετέπειτα ανταμοιβές. Επίσης, έχει μαθηματική αξία, καθώς εξασφαλίζει ότι η απόδοση είναι πάντα φραγμένη (για $\gamma < 1$).

Για $\gamma = 0$, ο πράκτορας είναι μυοπυκός, δηλαδή ενδιαφέρεται να εξασφαλίσει την καλύτερη ανταμοιβή σε κάθε βήμα, ανεξάρτητα αν αυτό σημαίνει ότι θα χάσει καλύτερες ανταμοιβές αργότερα, ενώ οσο η τιμή πηγαίνει προς το 1, ο πράκτορας βλέπει όλο και πιο μακριά.

Με βάση τα παραπάνω, μπορούμε πλεον να προσδιορίσουμε και πιο φορμαλιστικά την περιγραφή της πολιτικής του πράκτορα και της συνάρτησης αξίας. Οπως αναφέρθηκε ήδη η πολιτική είναι μια αντιστοίχηση μεταξύ καταστάσεων και πιθανοτήτων επιλογής πράξεων. Αν ένας πράκτορας ακολουθεί μια πολιτική π την χρονική στιγμή t, τότε $\pi(a|s)$ είναι η πιθανότητα ότι θα επιλεχθεί $A_t = a$, αν $S_t = s$.

Η συνάρτηση αξίας μιας κατάστασης s όταν ο πράκτορας ακολουθεί μια πολιτική π, με ένδειξη $v_\pi(s)$ είναι η αναμενόμενη απόδοση όταν ο πράκτορας ξεκινήσει από την κατάσταση s και ακολουθήσει την πολιτική π. Για ΜΔΑ, αυτή ορίζεται ως

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \;\middle|\; S_t = s \right] \tag{1.4}$$

για όλες τις καταστάσεις s του συνόλου $\mathcal{S}$. Το $\mathbb{E}_\pi[\cdot]$ είανι η αναμενόμενη τιμή μιας τυχαίας μεταβλητής δεδομένου ότι ο πράκτορας ακολουθεί πολιτική π και t είναι το χρονικό βήμα. Αυτή η συνάρτηση παραπάνω ορίζεται ως συνάρτηση αξίας-κατάστασης για την πολιτική π.

Μπορούμε να ορίσουμε και την συνάρτηση αξίας-ενέργειας για την πολιτική π, η οποία ορίζεται ως

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a \right] \tag{1.5}$$

η οποία προσδιορίζει την αξία του να κανει ο πράκτορας της ενέργεια a στην κατάσταση s υπο την πολιτική π, η οποία είναι η αναμενόμενη απόδοση που θα έχει ο πράκτορας αν στην κατάσταση s κάνει την ενέργεια a και μετά συνεχίσει με την πολιτική π.

Οι δύο συναρτήσεις αξίας, μπορούν να υπολογιστούν με βάση την εμπειρία που αποκτά ο πράκτορας κατά την μετάβαση του μεταξύ των καταστάσεων. Για παράδειγμα αν ακολουθεί την πολιτική π, τότε για κάθε κατάσταση, όσο πιο συχνά την επισκέφτεται, τόσο πιο κοντά θα φτάνει η μέση τιμή στην πραγματική αξία $v_\pi(s)$ της κατάστασης. Αντίστοια αν κρατάει μέσους όρους για κάθε πράξη σε κάθε κατάταση, θα φτάσει στην πραγματική αξία $q_\pi(s,a)$ της δράσης. Οι τεχνικές αυτές ονομάζονται Μόντε Κάρλο, γιατί βρίσκουν μέσους όρους πολλών τυχαίων δειγμάτων πραγματικών αποδόσεων. Ένας άλλος τρόπος είναι με χρήση συναρτήσεων που προσομοιώνουν τις συναρτήσεις αξιών, και είναι χρήσιμες όταν υπάρχει πολύ μεγάλος αριθμός καταστάσεων και η αποθήκευση όλων των τιμών δεν είναι δυνατή.

Στόχος της ΕΜ είναι η εύρεση βέλτιστων πολιτικών που θα επιφέρουν μεγάλη ανταμοιβή μακροπρόθεσμα. Σε πεπερασμένες ΜΔΑ η βέλτιστη απόφαση είναι πλήρως ορισμένη, και είναι η πολιτική που επιφέρει μεγαλύτερη ή ίση αξία με όλες τις υπόλοιπες πολιτικές. Αυτή η βέλτιστη πολιτική, δεν είναι να βρεθεί σε πραγματικά προβλήματα, γιατί

1. η δυναμική του περιβάλλοντος δεν είναι γνωστή ακριβώς,
2. δεν είναι υπολογιστικά εφικτό να υπολογιστεί η βέλτιστη πολιτική, πχ γιατί ο χώρος καταστάσεων είναι τεράστιος,
3. οι καταστάσεις δεν έχουν την Μαρκοβιανή ιδιότητα.

Σε αυτές τις περιπτώσεις, η μόνη μας επιλογή είναι είτε να δημιουργήσουμε προσεγγίσεις της βέλτιστης πολιτικής ή να χρησιμοποιήσουμε ευριστικές, ή και τα δύο.

1.4 Συμπεράσματα

Το πρόβλημα της Ενισχυτικής Μάθησης, μελετάει την έυρεση βέλτιστων πολιτικών για πράκτορες που κινούνται μέσα σε ένα περιβάλλον, πολιτικών δηλαδή που προσφέρουν την μέγιστη αξία στον πράκτορα. Επειδή πολλές η εύρεση της βέλτιστης πολιτικής δεν είναι πρακτικά δυνατή, χρησιμοποιούμε προσεγγίσεις και ευριστικές.

Στην δική μας εργασία, οι καταστάσεις του περιβάλλοντος είναι γνωστές και πεπερασμένες, και φραγμένες από τον αριθμό των θεμάτων που μπορεί να συζητήσει ο πράκτορας μας, οι οποίες δεν είναι πάρα πολλές. Επιπλέον, μπορούμε να θεωρήσουμε ότι οι καταστάσεις έχουν την Μαρκοβιανή ιδιότητα, καθώς η τρέχουσα κατάσταση μπορεί να περιλαμβάνει χαρακτηριστικά του διαλόγου του πράκτορα με τον χρήστη. Το πρόβλημα είναι ότι οι δυναμικές του συστήματος είναι πλήρως άγνωστες και μεταβλητές, καθώς δεν γνωρίζουμε καμία πληροφορία για τον χρήστη κατα την αρχή της συζήτησης για να μπορέσουμε να επιλέξουμε σωστά προτάσεις, και επιπλέον τα ενδιαφέροντα των χρηστών στο σύνολο μετακινούνται με την πάροδο του χρόνου. Επειδή ο συγκεκριμένος πράκτορας δεν αποκτά πολλές πληροοφορίες για το περιβάλλον, καθώς δεν υπάρχουν πολλοί χρήστες, θα πρέπει να απλοποιήσουμε το πρόβλημα και να μην ασχοληθούμε με το πλήρες πρόβλημα της ενισχυτικής μάθησης, αλλά με ένα πιο περιορισμένο, το πρόβλημα των bandits.

Κεφάλαιο 2

Bandits

2.1 Το πρόβλημα των bandits

Το πρόβλημα των ληστών (bandits), πήρε την ονομασία του από τα μηχανήματα του καζίνο, τους κουλοχέρηδες. Ο όρος one arm bandit προέρχεται από το γεγονός ότι οι κουλοχέρηδες έχουν ένα μοχλό-χέρι και σου ¨κλέβουν' τα χρήματα. Η ορολογία στα Ελληνικά δεν είναι ιδιαίτερα καθιερωμένη, οπότε και θα χρησιμοποιήσουμε την Αγγλική.

Το πρόβλημα των bandits είναι μια απλοποίηση του προβλήματος της Ενισχυτικής Μάθησης, και στην απλούστερη του μορφή το πρόβλημα δεν είναι προσεταιριστικό, δηλαδή κάθε κατάσταση θεωρείται ξεχωριστό γεγονός. Αυτό το γεγονός ότι οι αποφάσεις που κάνει ο πράκτορας στον ένα γύρο δεν επηρεάζουν τις ανταμοιβές και τις επιλογές του πράκτορα στους επόμενους γύρους, είναι και το πιο βασικό χαρακτηριστικό που μας ενδιαφέρει για να απλοποιήσουμε ελαφρώς και το δικό μας πρόβλημα. Ένα ακόμα σύνηθες χαρακτηριστικό των bandits είναι ότι ο πράκτορας μπορεί να παρατηρήσει τις ανταμοιβές του σε κάθε γύρο. Αν δεν μπορεί, τότε το πρόβλημα αυτό λέγεται πρόβλημα *μερικής παρακολούθησης* και δεν είναι κάτι που θα μας απασχολήσει περαιτέρω.

Παρόλο που το πρόβλημα μοίαζει φαινομενικά να είναι πολύ απλούστερο του πλήρους προβλήματος ΕΜ, η επίλυση του δεν είναι τόσο εύκολη. Σαν παράδειγμα [13], μπορούμε να σκεφτούμε ένα κουλοχέρη που έχει δύο μοχλούς, έναν δεξιό και έναν αριστερό, τους οποίους τραβώντας τους για 10 γύρους έχουμε τα παρακάτω αποτελέσματα.

Γύρος	1	2	3	4	5	6	7	8	9	10
Αριστερό	0		10	0		0				10
Δεξί		10			0		0	0	0	

Υπολογίζοντας την μέση ανταμοιβή που έχουμε από κάθε χέρι, μπορούμε να υπολογίσουμε ότι για το αριστερό χέρι αυτή είναι 4€, ενώ για το δεξί είναι 2€. Άρα το αριστερό χέρι είναι

φαινομενικά καλύτερο. Ποια θα ήταν η στρατηγική μας από εδώ και πέρα, αν είχαμε ακόμα 10 ακόμα προσπάθειες· Θα χρησιμοποιούσαμε μόνο το αριστερό χέρι για να εκμεταλλευτούμε αυτό που πιστεύουμε ότι είναι καλύτερο· Θα χρησιμοποιούσαμε το δεξί χέρι, για να εξερευνήσουμε και να μάθουμε αν έχουμε σωστή προσέγγιση της τιμής του· Το δίλημμα μεταξύ εκμετάλλευσης και εξερεύνησης είναι κεντρικό στα προβλήματα αυτά.

Η ορολογία που χρησιμοποιούμε στα προβλήματα bandits είναι η ίδια που χρησιμοποιήσαμε και στα προβλήματα ΕΜ, με την διαφορά ότι το πλήθος των γύρων που θα παίξει ο πράκτορας ονομάζονται **ορίζοντας**. Τα προβλήματα των bandits είναι προβλήματα που η δυναμική του περιβάλλοντος είναι άγνωστη, και έτσι ο πράκτορας πρέπει να την ανακαλύψει παίζοντας. Το μόνο που γνωρίζει ο πράκτορας για το περιβάλλον είναι ότι βρίσκεται σε μια οικογένεια περιβαλλόντων $\mathcal{E}$.

Κύρια μετρική της ποιότητας μιας πολιτικής είναι η μετάνοια (regret), η οποία εκφράζει την διαφορά μεταξύ των αναμενόμενων ανταμοιβών μιας πολιτικής π σε n γύρους, η οποία δεν είναι απαραίτητα αυτή που ακολούθησε ο πράκτορας, και των ανταμοιβών που πραγματικά πήρε ο πράκτορας σε n γύρους, σύμφωνα με την πολιτική που ακολούθησε. Είναι χρήσιμο να υπολογίζουμε την μετάνοια και σε σχέση με μια οικογένεια πολιτικών Π. Τότε η μετάνοια είναι η διαφορά της πολιτικής που ακολούθησε ο πράκτορας σε σχέση με την πολιτική η οποία έχει τις μεγαλύτερες ανταμοιβές μεταξύ των πολιτικών της οικογένειας Π (με άλλα λόγια την καλύτερη πολιτική). Συνήθως επιλέγουμε την οικογένεια Π, ώστε η βέλτιστη πολιτική για όλη την οικογένεια $\mathcal{E}$ να βρίσκεται μέσα στην οικογένεια πολιτικών Π. Ήδη διαισθητικά καταλαβαίνουμε ότι μεγάλη μετάνοια σημαίνει ότι ο πράκτορας δεν τα πάει καλά, ενώ μικρή ότι η πολιτική που ακολουθεί είναι κοντά στην βέλτιστη.

Για να μπορέσουμε να λύσουμε το πρόβλημα, συνήθως μειώνουμε τόσο την οικογένεια πρακτόρων, όσο και την οικογένεια των περιβαλλόντων, ώστε να περιέχει στοιχεία που έχουν συγκεκριμένες επιθυμητές ιδιότητες. Στόχος κάθε φορά είναι να δημιουργήσουμε αλγορίθμους που να πετυχαίνουν όσο καλύτερη μετάνοια είναι δυνατό.

Για παράδειγμα, μια εύκολη οικογένεια προβλημάτων είναι τα στοχαστικά, χρονικά αμετάβλητα προβλήματα bandits. Σε αυτή την οικογένεια προβλημάτων, το περιβάλλον παράγει ανταμοιβές με βάση μια πράξη, οι οποίες προέρχονται από μια κατανομή η οποία είναι σχετική στην πράξη αυτή και ανεξάρτητες από τις προηγούμενες πράξεις. Επίσης οι ανταμοιβές είναι χρονικά αμετάβλητες, είναι συναρτήσεις δηλαδή οι οποίες δεν έχουν ως παράμετρο τους τον χρόνο.

Από την άλλη, αν δεν θέλουμε να κάνουμε καμία υπόθεση για το περιβάλλον, θα μπορούσαμε να υποθέσουμε ότι το μόνο που ξέρουμε είναι ότι οι ανταμοιβές επιλέγονται χωρίς να υπάρχει γνώση των πράξεων του πράκτορα και απλά είναι στοιχεία σε ένα πεπερασμένο σύνολο. Ουσιαστικά αυτό είναι το πρόβλημα των ανταγωνιστικών (adversarial) bandits, όπου ουσιαστικά το περιβάλλον θεωρείται αντίπαλος. Ο αντίπαλος μπορεί να έχει πολύ μεγάλη ισχύ, ακόμα και την ικανότητα να δει τον κώδικα των αλγορίθμων και να διαλέξει ανταμοιβές αντίστοιχα. Παρ΄όλα αυτά το πρόβλημα αυτό δεν είναι πολύ δυσκολότερο από το στοχαστικό

πρόβλημα.

Ανάμεσα στα δύο αυτά άκρα, υπάρχουν πολλές επιλογές και υποθέσεις που μπορούμε να κάνουμε σχετικά με το περιβάλλον και τις πολιτικές.

2.2 Στοχαστικοί bandits

Πιο φορμαλιστικά ένα στοχαστικό σύστημα bandits είναι μια συλλογή απο κατανομές $v = (P_a : a \in \mathcal{A})$, όπου $\mathcal{A}$ είναι, όπως πάντα το σύνολο των δυνατών δράσεων. Το περιβάλλον και ο πράκτορας αλληλεπιδρούν διαδοχικά για n γύρους. Συνήθως το πλήθος των γύρων (ο ορίζοντας) είναι πεπερασμένος, αλλά πολλές η αλληλεπίδραση είναι αέναη. Σε κάθε γύρο $t \in \{1, \dots, n\}$ ο πράκτορας επιλέγει μια δράση $A_t \in \mathcal{A}$, η οποία τροφοδοτεί το περιβάλλον. Το περιβάλλον τότε επιλέγει μια ανταμοιβή $X_t \in \mathbb{R}$ από μια κατανομή P_{A_t} και επιστρέφει το X_t στον πράκτορα. Η αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος παράγει ένα μέτρο πιθανότητας στην αλληλουχία των αποτελεσμάτων $A_1, X_1, A_2, X_2, \dots, A_n, X_n$. Αυτή η αλληλουχία ικανοποιεί τις παρακάτω υποθέσεις:

1. Η δεσμευμένη πιθανότητα της ανταμοιβής X_t δεδομένου $A_1, X_1, \dots, A_{t-1}, X_{t-1}, A_t$ είναι P_{A_t} που διαισθητικά σημαίνει ότι το περιβάλλον παίρνει ένα δείγμα X_t από την κατανομή P_{A_t} στον γύρο t.
2. Ο δεσμευμένος κανόνας της δράσης A_t δεδομένων $A_1, X_1, \dots A_{t-1}, X_{t-1}$ είναι $\pi(\cdot|A_1, X_1, \dots, A_{t-1}, X_{t-1})$ όπου $\pi_1, \pi_2, \dots$ είναι η αλληλουχία από Μαρκοβιανούς πυρήνες που περιγράφουν τον πράκτορα. Διαισθητικά αυτό σημαίνει ότι ο πράκτορας δεν μπορεί να χρησιμοποιήσει παρατηρήσεις από το μέλλον σε τρέχουσες αποφάσεις.

Ο στόχος του πράκτορα είναι να μεγιστοποιήσει την συνολική ανταμοιβή $S_n = \sum_{t=1} nX_t$ (η οποία διαφέρει ελαφρά από την ανταμοιβή του προβλήματος EM, καθώς εδώ $\gamma = 1$). Η συνολική ανταμοιβή είναι μια τυχαία ποσότητα η οποία εξαρτάται από τις πράξεις του πράκτορα και τις ανταμοιβές που πήρε από το περιβάλλον.

Αν έχουμε την πολιτική $v = (P_a : a \in \mathcal{A})$, τότε μπορούμε να ορίσουμε το μέσο όρο κάθε χεριού

$$\mu_\alpha(v) = \int_{-\infty}^{\infty} x dP_a(x)$$

Τότε μπορούμε να ορίσουμε το $\mu^*(v) = \max_{a \in \mathcal{A}} \mu_a(v)$ ο μέγιστος μέσος όρος των χεριών.

Τότε η μετάνοια της πολιτικής π σε ένα πρόβλημα bandit είναι

$$R_n(\pi, v) = n\mu^*(v) - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right] \tag{2.1}$$

όπου η αναμενόμενη τιμή υπολογίζεται με βάση την πιθανότητα των αποτελεσμάτων που δημιουργούνται από την αλληλεπίδραση του π και του v.

Όλες οι πολιτικές βασίζονται στην παρατήρηση ότι για να μειώσουμε την μετάνοια, ο αλγόριθμος πρέπει να ανακαλύψει την δράση/χέρι με τον μεγαλύτερο μέσο όρο. Συνήθως αυτό σημαίνει ότι ο πράκτορας πρέπει να παίξει κάθε χέρι κάποιον αριθμό φορών, ώστε να δημιουργήσει μια εκτίμηση της μέσης τιμής του χεριού, και στην συνέχεια να παίξει το χέρι με την μεγαλύτερη τιμή. Έτσι το πρόβλημα μπορεί να συνοψιστεί ως την προσπάθεια να ανακαλύψει ο πράκτορας πόσο συχνά πρέπει να παίξει κάθε χέρι, ώστε να μπορεί με στατιστική βεβαιότητα να πει ότι έχει βρει το βέλτιστο χέρι.

2.3 Ανταγωνιστικοί Bandits

Το πλαίσιο των ανταγωνιστικών bandits έχει τις ρίζες του στην θεωρία παιγνίων. Ένα παράδειγμα ενός τέτοιου προβλήματος είναι το εξής παιχνίδι. Παίζουμε μια ένα φίλο μας ένα απλό παιχνίδι με bandits, όπου ο ορίζοντας είναι $n = 1$ και έχουμε 2 δράσεις. Το παιχνίδι έχει την ακόλουθη μορφή:

- Λέμε στον φίλο μας την στρατηγική με βάση την οποία θα επιλέξουμε την δράση.
- Ο φίλος μας διαλέγει κρυφά ανταμοιβές $x_1 \in \{0, 1\}$ και $x_2 \in \{0, 1\}$.
- Εφαρμόζουμε την στρατηγική που επιλέξαμε $A \in \{1, 2\}$ και παίρνουμε ανταμοιβή x_A.
- Η μετάνοια είναι $R = \max x_1, x_2 - x_A$

Προφανώς αν ο φίλος μας επιλέξει και τις δύο ανταμοιβές να είναι 0, τότε η μετάνοια θα είναι πάντα 0. Ο τρόπος για να είναι η στρατηγική επιτυχημένη είναι η τυχαιότητας στις επιλογές μας. Έτσι παρόλο που ο αντίπαλος γνωρίζει την στρατηγική μας, δεν γνωρίζει ακριβώς τις επιλογές που θα κάνουμε. Για παράδειγμα, μπορούμε να πούμε στον φίλο μας, ¨Θα επιλέξω την κίνηση $A = 1$ με πιθανότητα 1/2' και η αναμενόμενη μετάνοια γίνεται $R = 1/2$. Οσο μεγαλώνει ο ορίζοντας, το πλεονέκτημα του αντιπάλου όλο και μειώνεται.

Πιο φορμαλιστικά, αν $k > 1$ ο αριθμός των χεριών, τότε ένα πρόβλημα ανταγωνιστικού bandit k-χεριών είναι μια αυθαίρετη σειρά από διανύσματα ανταμοιβών $(x_t)_{t=1}^n$, όπου $x_t \in [0, 1]^k$. Σε κάθε γύρο ο πράκτορας διαλέγει μια κατανομή πράξεων $P_t \in P_{k-1}$. Τότε η δράση $A_t \in [k]$ είναι ένα δείγμα από την κατανομή P_t, και ο πράκτορας λαμβάνει ανταμοιβή x_{tA_t}.

Η πολιτική σε αυτή την περίπτωση είναι μια συνάρτηση $\pi : ([k] \times [0, 1])^* \to \mathcal{P}_{k-1}$, η οποία χαρτογραφεί ακολουθίες της ιστορίας σε κατανομές πάνω σε πράξεις. Η επίδοσης της πολιτικής π σε ένα περιβάλλον x μετριέται από την αναμενόμενη μετάνοια, η οποία είναι η αναμενόμενη απώλεια σε κέρδος της πολιτικής π σε σχέση με την καλύτερη πολιτική που επιλέγει ένα χέρι κάθε φορά.

$$R_n(\pi, x) = \max_{i \in [k]} \sum_{t=1}^{n} x_{t_i} - \mathbb{E}\left[\sum_{t=1}^{n} x_{tA_t}\right] \tag{2.2}$$

όπου η αναμενόμενη τιμή είναι πάνω στην τυχαιοτητα των πράξεων του πράκτορα.

Η μετάνοια χειρότερης περίπτωσης σε όλα τα περιβάλλοντα είναι

$$R_n^*(\pi) = \sup_{x \in [0,1]^{n \times k}} R_n(\pi, x)$$

Για να φτιάξουμε πολιτικές που είναι υπο-γραμμικές στο n στην χειρότερη περίπτωση, δηλαδή πολιτικές π που ισχύει

$$\lim_{n \to \infty} \frac{R_n^*(\pi)}{n} = 0$$

θα πρέπει να χρησιμοποιήσουμε πολιτικές με τυχαιότητα.

2.4 Γνωστοί αλγόριθμοι

Παρακάτω παρουσιάζονται κάποιοι από τους πιο γνωστούς αλγορίθμους bandits. Οι ε-άπληστοι (ε-greedy), Ανώτατο Όριο Εμπιστοσύνης (Upper Confidence Bound - UCB) και η δειγματοληψία Thompson αποτελούν λύσεις στο πρόβλημα των στοχαστικών bandits, ενώ ο EXP3 αποτελεί λύση στο πρόβλημα των ανταγωνιστικών bandits.

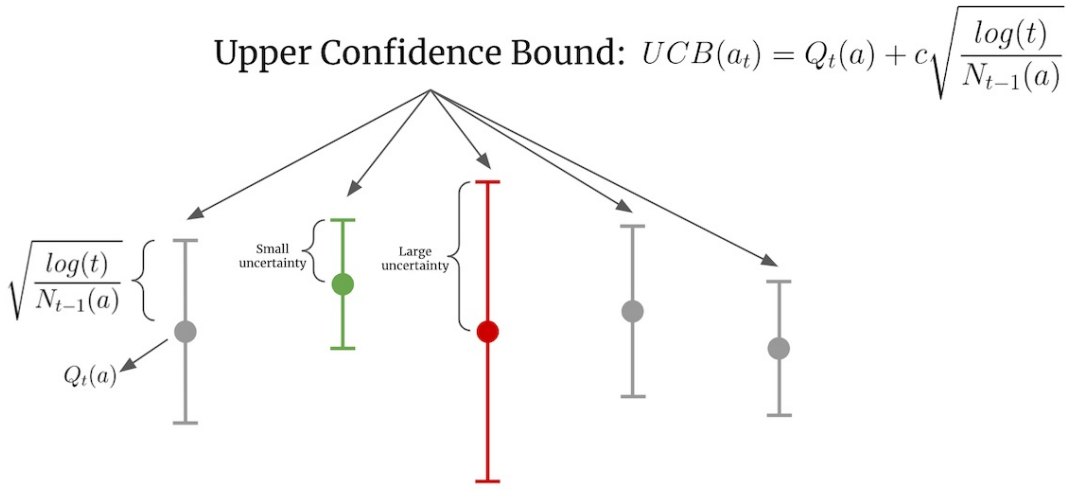
2.4.1 Αλγόριθμος ε-greedy

Ο αλγόριθμος ε-greedy είναι ο απλούστερος αλγόριθμος και ίσως η πιο προφανής λύση του διλήμματος μεταξύ εξερεύνησης και εκμετάλλευσης. Η πολιτική εξερευνά ένα τυχαίο χέρι με πιθανότητα ε, ενώ με πιθανότητα $1-\varepsilon$, η πολιτική εκμεταλλεύεται την λύση με την μεγαλύτερη ανταμοιβή κατά μέση τιμή. Στην κλασική έκδοση του αλγορίθμου το ε είναι σταθερά, αλλά αυτό δεν είναι απαραίτητο. Αντίθετα βγάζει νόημα το ε να εξαρτάται από τις επαναλήψεις (γραμμική μείωση, εκθετική μείωση, εξερεύνηση με πιθανότητα ε για κάποιες επαναλήψεις και εκμετάλλευση με πιθανότητα $1-\varepsilon$ και καμία εξερεύνηση αργότερα. Έτσι η επόμενη κίνηση A_t επιλέγεται απο τον τύπο.

$$A_t = \begin{cases} randint(1, k), & \text{if } n \leq \varepsilon \\ \underset{a}{argmax}\ Q_{t-1}(a), & \text{otherwise} \end{cases}$$

όπου το n είναι μια τυχαία μεταβλητή η οποία προέρχεται από μια ομοιόμορφη κατανομή ανάμεσα στο 0 και το 1. Η $randint$ είναι μια συνάρτηση που επιστρέφει ένα συγκεκριμένο ακέραιο μέσα στο δοθέν εύρος, k είναι το πλήθος των χεριών, $Q_{t-1}(a)$ είναι η αναμενόμενη μέση τιμή του χεριού του $a-$οστού χεριού την χρονική στιγμή $t-1$.

Αυτή η πολιτική είναι απλή και υπάρχουν πολιτικές που επιφέρουν καλύτερη μετάνοια, αφού υπάρχουν πιο έξυπνοι τρόποι εξερεύνησης σε σχέση με την τυχαία επιλογή. Αφού ο αλγόριθμος έχει τρέξει για κάποιους γύρους, μπορούμε να δούμε ήδη ότι κάποια από τα μπράστα έχουν κακή απόδοση, και δεν υπάρχει ανάγκη περαιτέρω εξερεύνησης τους, οπότε αυτή η εξερεύνηση θα μπορούσε να χρησιμεύσει για χέρια που έχουν καλύτερες πιθανότητες να είναι το βέλτιστο χέρι.



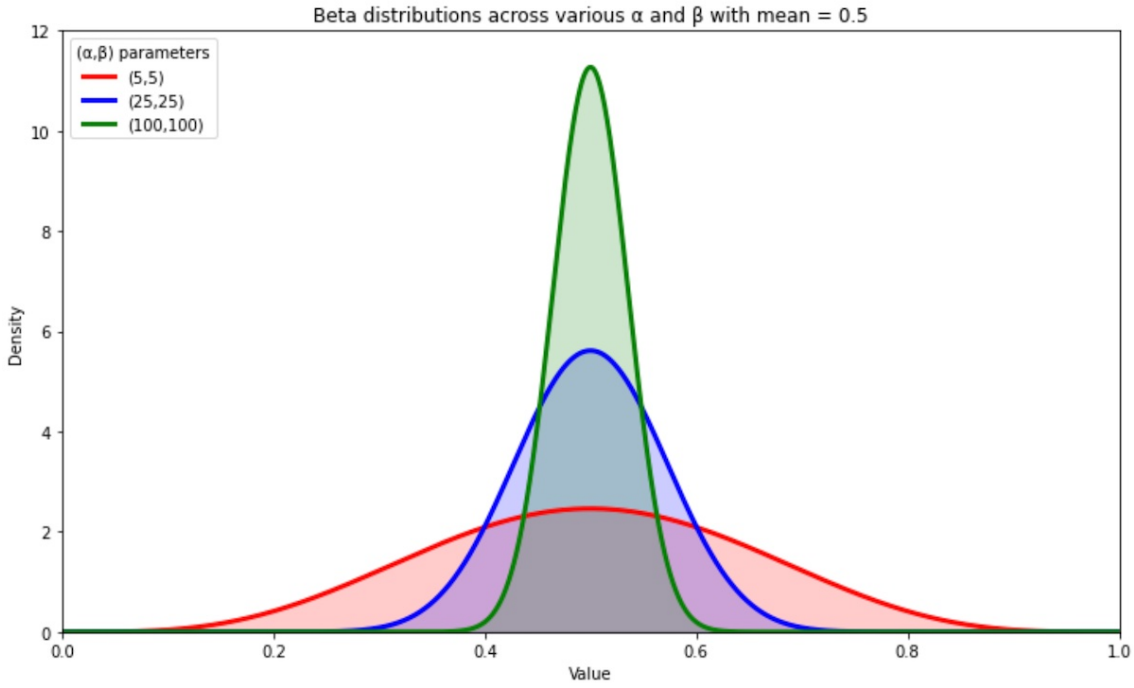
Σχήμα 2.1: Οπτικοποίηση UCB [17] διάφορων χεριών

2.4.2 Upper Confidence Bound (UCB)

Η βασική ιδέα του αλγορίθμου Μέγιστου Ορίου Εμπιστοσύνης (Upper Confidence Bound - UCB) είναι η επιλογή πάντα του χεριού με το υψηλότερο μέγιστο όριο. Μπορεί να περιγραφή σαν αισιοδοξία στην αντιμετώπιση αβεβαιότητας. Το προβλεπόμενο μέγιστο όριο αποτελείται από δύο στοιχεία: την προβλεπόμενη μέγιστη ανταμοιβή και την αβεβαιότητα, όπως φαίνονται στην εξίσωση.

$$A_t = \underset{a}{argmax} \left[Q_{t-1}(a) + \sqrt{\frac{\log(t-1)}{N_{t-1}(a)}} \right] \tag{2.3}$$

όπου $Q_{t-1}(a)$ είναι η προβλεπόμενη μέση ανταμοιβή του a-οστού χεριού την χρονική στιγμή $t-1$, $t-1$ είναι το πλήθος των χεριών που έχουν τραβηχθεί μέχρι τώρα (ή ο αριθμός των βημάτων γενικότερα) και N_{t-1} είναι το πλήθος των φορών που το a-οστό χέρι έχει τραβηχθεί. Έτσι χέρια με μεγαλύτερη μέση ανταμοιβή έχουν μεγαλύτερη τιμή μεγίστου ορίου. Χέρια που δεν έχουν εξερευνηθεί, τείνουν να έχουν καλύτερα σκορ λόγω εκτιμήσεων αβεβαιότητας. Αυτό θα επιφέρει μικρότερη μετάνοια σε σχέση με τον ε-άπληστο αλγόριθμο, καθώς μικρότερο ποσό εξερεύνησης θα καταναλωθεί σε εμφανώς μη-βέλτιστα χέρια. Οπτικά τα παραπάνω φαίνονται στο Σχήμα 2.1, όπου έχουμε ένα πράσινο χέρι που έχει επιλεχθεί πολλές φορές και ένα κόκκινο που έχει επιλεχθεί λίγες. Όπως βλέπουμε το άνω όριο εμπιστοσυνης του κόκκινου είναι το ψηλότερο σε αυτό το βήμα, οπότε αυτό είναι το χέρι που θα επιλεχθεί, το διάστημα εμπιστοσύνης του θα μικρύνει και το κέντρο θα μετακινηθεί ανάλογα με την ανταμοιβή.



Σχήμα 2.2: Η κατανομή Βήτα στενεύει όσο τα α και β μεγαλώνουν

2.4.3 Δειγματοληψία Thompson

Η δειγματοληψία Thompson χτίζει μια κατανομή πιθανότητας βασισμένη σε ιστορικές ανταμοιβές και μετά δειγματοληπτεί από την κατανομή καθε δράσης για την επιλογή αυτής που επιφέρει την μέγιστη αναμενόμενη ανταμοιβή. Στην απλή περίπτωση που η ανταμοιβή είναι δυαδική (0 ή 1) και άρα θέλουμε να υπολογίσουμε την πιθανότητα να υπάρχει ανταμοιβή, χρησιμοποιειται η κατανομή Βήτα για την μοντελοποίηση των κατανομών των ανταμοιβών του κάθε χεριού. Η κατανομή Βήτα παίρνει δύο παραμέτρους, τα α και β, όπου α είναι οι φορές που η ανταμοιβή είναι 1 και β είναι η φορές που η ανταμοιβή ήταν 0. Η μέση τιμή της κατανομής είναι $\frac{\alpha}{\alpha+\beta}$, το οποίο αντιστοιχεί στο κλάσμα των επιτυχιών προς το σύνολο των προσπαθειών. Για την επιλογή μιας δράσης, δειγματοληπτούμε από την κατανομή Βήτα κάθε χεριού και επιλέγουμε τον χέρι με την υψηλότερη δειγματολημμένη τιμή [10].

Η κατανομή Βήτα για διάφορα α και β φαίνεται στο Σχήμα 2.2. Όσο αυξάνεται το πλήθων των α και β, η κατανομή στενεύει και άρα βρίσκεται πιο κοντά στην πιθανότητα που έχει το κάθε χέρι να επιφέρει ανταμοιβή. Έτσι χέρια που έχουν δοκιμαστεί λιγότερο και έχουν πιο διευρυμένη κατανομή, έχουν πιθανότητες να δοκιμαστούν ως εξερεύνηση για να εξεταστεί αν επιφέρουν καλύτερες ανταμοιβές.

Η δειγματοληψία Thompson δουλεύει και σε περιπτώσεις που η ανταμοιβή δεν είναι δυα-

δική, απλά η κατανομή που χρησιμοποιειται δεν είναι η Βήτα, αλλά κάποια άλλη.

2.4.4 Ο αλγόριθμος EXP3

Το EXP3 σημαίνει Exponential-weight algorithm for Exploration and Exploitation, δηλαδή Αλγόριθμος εκθετικών βαρών για αναζήτηση και εκμετάλλευση. Ο τρόπος που δουλεύει είναι διατηρώντας μια λίστα με τα βάρη κάθε δράσης και χρησιμοποιώντας τα για να επιλέξει τυχαία ποια δράση να κάνει μετά. Τέλος, τα αντίστοιχα βάρη αυξάνονται/μειώνονται όταν η ανταμοιβή είναι καλή/κακή. Επίσης υπάρχει ο παράγοντας γ, ο οποίος ορίζει την θέληση να επιλέξουμε μια δράση με ομοιόμορφη τυχαιότητα. Έτσι, αν $\gamma = 1$, τα βάρη δεν έχουν καμία επίδραση στις επιλογές σε κάθε βήμα.

Ο αλγόριθμος μπορεί να περιγραφεί ως:

1. Δοθέντος $\gamma \in [0, 1]$, αρχικοποιούμε τα βάρη $w_i(1) = 1$ για $\iota = 1, \ldots, K$, όπου K είναι τα χέρια.
2. Για κάθε γύρο t:
   (α′) Θέτουμε $p_i(t) = (1 - \gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$ για κάθε i.
   (β′) Επιλέγουμε την επόμενη δράση i_t τυχαία με βάση την κατανομή του $p_i(t)$.
   (γ′) Παρατηρούμε την ανταμοιβή $x_{i_t}(t)$.
   (δ′) Ορίζουμε την αναμενόμενη ανταμοιβή $\hat{x}_{i_t}(t)$ να είναι $x_{i_t}(t)/p_{i_t}(t)$. Αυτό το βήμα εξασφαλίζει ότι η δεσμευμένη προσδοκία της αναμενόμενης ανταμοιβής είναι η πραγματική ανταμοιβή.
   (ε′) Θέτουμε $w_{i_t}(t + 1) = w_{i_t}(t)e^{\gamma \hat{x}_{i_t}(t)/K}$
   (ϛ′) Θέτουμε όλα τα άλλα $w_j(t + 1) = w_j(t)$.

2.5 Contextual Bandits

Η περίπτωση bandits που μας ενδιαφέρει είναι αυτή στην οποία ο αλγόριθμος έχει πρόσβαση σε πληροφορίες σχετικά με το συγκείμενο του περιβάλλοντος, τις οποίες θα μπορούσε να χρησιμοποιήσει για να πάρει καλύτερες αποφάσεις. Το πρόβλημα και η μετρική (η μετάνοια) που μελετήσαμε νωρίτερα δεν χρησιμοποιούσαν τέτοια δεδομένα και προσπαθούσαν να επιλέξουν την καλύτερη κίνηση. Το πρόβλημα αυτό στην μορφή που περιγράφουμε εδώ, μελετήθηκε από τον John Langford και τον Tong Zhang στο [3], καθώς σε προβλήματα στον πραγματικό κόσμο πάντα υπάρχουν επιπλέον πληροφορίες που μπορεί να χρησιμοποιήσει ο πράκτορας για να πάρει καλύτερες αποφάσεις. Συγκεκριμένα το πρόβλημα που προσπαθούσαν να λύσουν είναι η αντιστοίχηση διαφημισεων σε περιεχόμενο ιστοσελίδων στο ίντερνετ.

Για να μελετήσουμε αυτή την νέα περίπτωση θα χρειαστεί να επεκτείνουμε το πλαίσιο στο οποίο εργαζόμαστε, καθώς για τον ορισμό της μετάνοιας, ώστε μπορέσουμε να μοντελοποιήσουμε αυτα τα προβλήματα, τα οποία περιέχουν πληροφορίες συγκειμένου. Είναι σημαντικό να έχουμε υπ΄όψιν ότι όταν σχεδιάζουμε μια καινούρια μετρική έχουμε να συμβιβαστούμε μεταξύ της μεροληψίας και της διακύμανσης bias-variance trade-off. Μεροληψίας από την άποψη ότι δεν θέλουμε να βρούμε μια κακή μετρική με την οποία να συγκριθούμε, γιατί τότε κάθε αλγόριθμος που θα έχει παρόμοια απόδοση με την μετρική θα έχει κακή απόδοση. Από την άλλη, ο συναγωνισμός με μια καλύτερη μετρική μπορεί να είναι πολύ δύσκολος από την προοπτική της μάθησης και αυτή η τιμωρία μπορεί να υπερτερεί των πλεονεκτημάτων.

Αν προσεγγίσουμε τους contextual bandits με χρήση των ιδεών από τους ανταγωνιστικούς bandits, τότε το πρόβλημα παίρνει την παρακάτω μορφή:

1. Ο αντίπαλος κρυφά διαλέγει ανταμοιβές $(x_t)_{t=1}^n$ με $x_t \in [0,1]^k$
2. Ο αντίπαλος κρυφά διαλέγει συγκείμενο $(c_t)_{t=1}^n$ με $c_t \in \mathcal{C}$, όπου $\mathcal{C}$ είναι το σταθερό σύνολο των πιθανών συγκειμένων.
3. Για τους γύρους $t = 1, 2, \ldots, n$:
   (α΄) Ο πράκτορας παρατηρεί συγκείμενο $c_t \in \mathcal{C}$
   (β΄) Ο πράκτορας διαλέγει κατανομή $P_t \in \mathcal{P}_{k-1}$ και παίρνει δείγμα A_t από το P_t
   (γ΄) Ο πράκτορας παρατηρεί ανταμοιβή $X_t = x_{tA_t}$

Το πλαίσιο των contextual bandits μας προσφέρει το εργαλείο για να περιγράψουμε το πρόβλημα των συστάσεων που θέλουμε να λύσουμε. Μετά την εισαγωγή του προβλήματος σαν μέθοδο για την επίλυση τέτοιου είδους προβλημάτων το 2007, έχει υπάρξει σημαντική έρευνα, αλλά και χρήση στην βιομηχανία τέτοιων μεθόδων, ειδικά σε περιπτώσεις που το πλήθος των αντικειμένων αλλάζει δυναμικά, πχ νέα άρθρα προστίθενται κάθε μέρα, ενώ τα παλιά άρθρα είναι πλεον λιγότερο σημαντικά. Έτσι εταιρίες όπως η Microsoft, το LinkedIn, το Netflix και άλλες χρησιμοποιούν τέτοιες μεθόδους για να προτείνουν άρθρα, διαφημίσεις ή ταινίες αντίστοιχα. Το Netflix χρησιμοποιεί επίσης contextual bandits για να επιλέξει την εικόνα που δείχνει για κάθε ταινία [9]. Στα επόμενα μέρη της διπλωματικής, δεν αναλύσουμε περισσότερο τις ιδιότητες των αλγορίθμων αυτών, ουτε θα προχωρήσουμε σε αναλύσεις της μετάνοιας τους, αλλά θα τους χρησιμοποιήσουμε μόνο βάση της ανάλυσης στις εργασίες που εισήχθηκαν.

Κεφάλαιο 3

Διάλογος και Συστάσεις

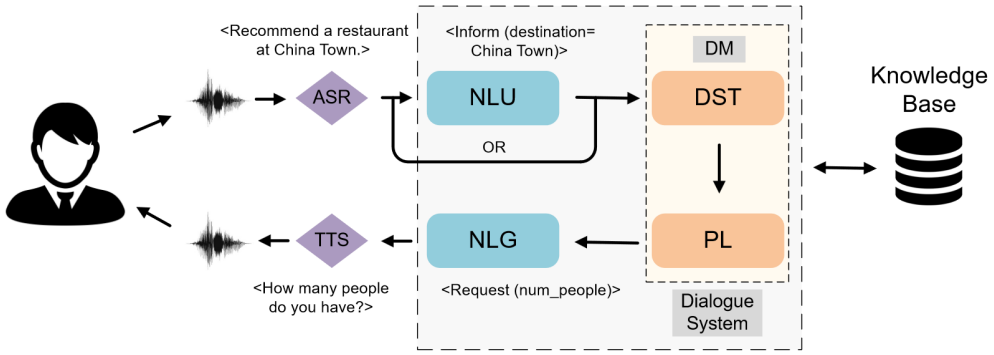
Σε αυτό το κεφάλαιο θα περιγράψουμε πολύ περιληπτικά την χρήση της ενισχυτικής μάθησης σε διαλογικά συστήματα, ώστε να μπορέσουμε να το συνδέσουμε με το πρόβλημα μας, και μετά θα περιγράψουμε την χρήση ενισχυτικής μάθησης σε προβλήματα συστάσεων, που είναι και πιο κοντά στο πεδίο του προβλήματος μας.

3.1 Διάλογικά Συστήματα

Η ιδέα της δημιουργίας ενός πράκτορα που θα μπορεί να απαντήσει σε ανθρώπινες ερωτήσεις, ξεκίνησε από το σύγγραμα του Alan Turing, Computing Machinery and Intelligence [1]. Τα πρώτα μοντέλα που δημιουργήθηκαν βασίζονταν σε κανόνες, όπου αναγνώριζαν κάποιες λέξεις-κλειδία στο κείμενο και ανάλογα με αυτές απαντούσαν στον χρήστη. Το πρόβλημα αυτών των συστημάτων ήταν ότι ήταν πολύ δύσκολο να επεκταθούν, καθώς και να γενικεύσουν, αφού η δημιουργοί πρέπει να προσθέσουν χειροκίνητα τους επιπλέον κανόνες. Τα τελευταία χρόνια, η δημιουργία διαλογικών συστημάτων γίνεται ολοένα και περισσότερο με χρήση βαθειάς μηχανικής μάθησης, παρόλο που η χρήση κανόνων είναι ακόμα βολική σε κάποιες περιπτώσεις.

Τα διαλογικά συστήματα συνήθως χωρίζονται σε δύο κατηγορίες ανάλογα με τον σκοπό τους:

1. Συγκεκριμένου σκοπού (task-oriented systems). Τα διαλογικά αυτά συστήματα έχουν ως στόχο να βοηθήσουν τον χρήστη να πετύχει συγκεκριμένους στόχους, ιδανικά σε όσο λιγότερους γύρους διαλόγου γίνεται. Για παράδειγμα τέτοια συστήματα είναι συστήματα μέσω των οποίων ο χρήστης μπορεί να κλείσει εισητήρια, ή να λάβει υποστήριξη σχετικά με ένα πρόβλημα του.
2. Ανοιχτού σκοπού (open-domain systems). Τα διαλογικά συστήματα αυτά δεν έχουν κάποιο σκοπό, αλλά εστιάζουν στο να δώσουν ρεαλιστικές απαντήσεις σε συζητήσεις με τον χρήστη.



Σχήμα 3.1: Σύστημα συγκεκριμένου σκοπού [15]

Συχνά τα συστήματα συγκεκριμένου σκοπού οργανώνονται σε μια αλληλουχία μερών όπως αυτή που φαίνεται στο Σχήμα 3.1:

- **Το κομμάτι της κατανόησης της εισόδου του χρήστη.** Αυτό το κομμάτι είνα υπευθυνο για την ταξινόμηση των διάφορων λέξεων σε μέρη του λόγου, την αναγνώριση ονομάτων, αλλά και την αναγνώριση της πρόθεσης του χρήστη με βάση το τι είπε. Κάποια συστήματα δεν χρησιμοποιούν αυτό το κομμάτι και χρησιμοποιούν το ίδιο το μήνυμα του χρήστη ως είσοδο στο επόμενο κομμάτι, όπως στο Σχήμα 3.1. Αυτό συμβαίνει για να μειώσουν την επίδραση του λάθους του πρώτου αυτού του κομματιού και την μεταφορά λαθών στα μετέπειτα κομμάτια.
- **Το κομμάτι της διαχείρισης της κατάστασης του διαλόγου,** το οποίο ρυθμίζει τις καταστάσεις του διαλόγου με βάση την τρέχουσα είσοδο και την ιστορία του διαλόγου. Η κατάσταση του διαλόγου περιέχει σχετικές δράσεις του χρήστη και ζευγάρια θέσης-τιμής.
- **Το κομμάτι της εκμάθησης της πολιτικής του διαλόγου**, το οποίο με βάση τις καταστάσεις του διαλόγου που παίρνει από το προηγούμενο κομμάτι, επιλέγει την επόμενη δράση του διαλογικού πράκτορα.
- **Το κομμάτι της παραγωγής της απάντησης του συστήματος**, το οποίο μετατρέπει τις δράσεις που επιλέχθηκαν από το προηγούμενο κομμάτι σε φυσική γλώσσα, η οποία θα επιστραφεί στον χρήστη.

Άλλες φορές προτιμάται ένα σύστημα που υλοποιεί όλες τις παραπάνω λειτουργίες από άκρη σε άκρη, το οποίο μπορεί να πετύχει καλύτερη βελτιστοποίηση, καθώς δεν υπάρχει μεταφορά σφάλματος μεταξύ των διάφορων κομματιών.

3.1.1 Ενισχυτική Μάθηση και Διαλογικά Συστήματα

Όσον αφορά την χρήση ενισχυτικής μάθησης, σε συστήματα συγκεκριμένου σκοπού είναι στην διαχείρηση του διαλόγου. Πιο συγκεκριμένα δύο σύνηθεις χρήσεις είναι για την παρακολούθηση της κατάστασης του διαλόγου και την εκμάθηση της πολιτικής. Η δεύτερη χρήση είναι ιδιαίτερα σύνηθης και αρκετά επιτυχημένη, καθώς περιγράφεται ακριβώς από ένα πρόβλημα ΕΜ (πχ [12]).

Σε συστήματα ανοιχτού σκοπού, η χρήση της ενισχυτικής μάθησης είναι κυρίως η επιλογή απαντήσεων παρά η παραγωγή τους, καθώς τα παραγωγικά generative συστήματα είναι πολύ καλύτερα στην παραγωγή λόγου.

Για παράδειγμα, σε μια από τις πρώτες δουλείες οι οποίες χρησιμοποιήσαν ΕΜ σε διάλογο [7], οι συγγραφείς προσπάθησαν να ενώσουν τις ιδέες από τα seq2seq μοντέλα και την ΕΜ, ώστε να δημιουργήσουν συστήματα τα οποία επιστρέφουν καλύτερες απαντήσεις. Έτσι δημιούργησαν μια μετρική ανταμοιβής η οποία αξιολογούσε την ποιότητα των απαντήσεων. Αρχικά εκπαίδευσαν ένα seq2seq μοντέλο με επιβλεπόμενη μάθεση, και μετά με βάση αυτό προσομοιώσαν διαλόγους μεταξύ δυο πρακτόρων, ξεκινώντας από μια πρόταση από το σύνολο εκπαίδευσης.

Ενα από τα πιο διάσημα σύχρονα παραδείγματα είναι η χρήση του στην εκπαίδευση του InstructGPT. Στο [16], οι ερευνητές έκαναν fine-tune το ΓΠΤ-3 με χρήση ενισχυτικής μάθησης και ανατροφοδότησης από ανθρώπους. Συγκεκριμένα, κατα το fine-tuning, το μοντέλο ¨ρώταγε' τους χρήστες ποιά από τις απαντήσεις θεωρούσαν καλύτερη σε σχέση με ένα ερώτημα, οι χρήστες ταξινομούσαν τις απαντήσεις από καλύτερη προς χειρότερη, και με βάση αυτό εκπαιδεύτηκε ένα μοντέλο ανταμοιβών. Έπειτα, για την εκπαίδευση της πολιτικής των απαντήσεων του συστήματος, ενα ερώτημα επιλεγόταν από το σύνολο των δεδομένων, η πολιτική του συστήματος παράγει μια έξοδο και το μοντέλο ανταμοιβών επιλέγει την ανταμοιβή για αυτή την έξοδο. Με βάση αυτό ανανεώνεται η ανταμοβή της πολιτικής.

3.2 Συστήματα Συστάσεων

Ένα σύστημα συστάσεων αποτελείται από εργαλεία και αλγορίθμους οι οποίοι αναπτύχθηκαν με την ιδέα να βοηθήσουν τους χρήστες να βρουν αντικείμενα που τους ενδιαφέρουν. Σε μια γενική μορφή, ο στόχος είναι η δημιουργία του προφίλ των χρηστών βασισμένη στην ανατροφοδότηση μεταξύ συστήματος και χρήστη και η σύσταση αντικειμένων που να ταιριάζουν στο προφίλ αυτό. Το πρόβλημα απαντάται σε πολλούς κλάδους όπως της υγείας, της διασκέδασης, των νέων, κλπ.

Παραδοσιακά το πρόβλημα των συστάσεων θεωρούνταν ένα πρόβλημα ταξινόμησης ή πρόβλεψης, αλλά πλεον ο ακαδημαικός κόσμος συμφωνεί ότι η αλληλεπίδραση μεταξύ χρήστη και συστήματος μοντελοποιήται καλύτερα ως ενα πρόβλημα αποφάσεων με διαδοχικά βήματα [14]. Έτσι μπορεί να περιγραφεί από μια Μαρκοβιανή Διαδικασία Αποφάσεων και να λυθεί με χρήση ενισχυτικής μάθησης.

Πρωτού περάσουμε σε τεχνικές με χρήση ενισχυτικής μάθησης, είναι σημαντικό να γνωρίσουμε περιληπτικά τους κλασικούς αλγορίθμους. Αυτοί είναι:

- *Συνεργατικό φιλτράρισμα*: Η ιδέα της μεθόδου είναι η ομαδοποίηση του χρήστη (ή των αντικειμένων) σε ομάδες με παρόμοια χαρακτηριστικά. Όταν το φιλτράρισμα γίνεται με βάση τον χρήστη, οι προτάσεις γίνονται με βάση τις παρόμοιες προτιμήσεις διάφορων χρηστών. Από την άλλη, όταν αναφερόμαστε σε φιλτράρισμα με βάση τα αντικείμενα, οι προτάσεις γίνονται με βάση τα αντικείμενα τα οποία σχετίζονται με αυτά που ο χρήστης έχει ήδη αλληλεπιδράσει. Η μέθοδος αυτή μπορεί να χωριστεί σε δύο προσεγγίσεις: με βάση την μνήμη, όπου ουσιαστικά για κάθε χρήστη/αντικείμενο γίνεται μια σύγκριση ομοιότητας (πχ ομοιότητα συνημιτόνου) με τους υπόλοιπους και μετά με χρήση k-κοντινότερων γειτόνων, ή με βάση κάποιο μοντέλο, όπου ουσιαστικά δημιουργήται ένα μοντέλο που εκπαιδεύεται με τεχνικής μηχανικής μάθησης να βρίσκει ομοιότητες μεταξύ χρήστες [6].

- *Παραγοντοποίηση πινάκων*: Αυτή η μέθοδος χρησιμοποιήται και ξεχωριστά και ως μέρος του φιλτραρίσματος. Η ιδέα είναι η αναπαράστααση του χρήστη και των αντικειμένων ως αντικείμενα σε ένα χώρο λίγων διαστάσεων. Έπειτη η συμβατότητα χρήστη και προιόντος υπολογίζεται είτε με χρήση εσωτερικού γινομένου, ή με χρήση κάποιου νευρωνικού δικτύου αν οι σχέσει είναι μη γραμμικές.

  Οι κλασικές μέθοδοι, έχουν διάφορα προβλήματα, όπως το ότι το σύστημα δεν μπορεί να προσφέρει χρήσιμες συστάσεις σε ένα νέο χρήστη, ή οταν προστίθεται ένα νέο αντικείμενο (cold-start), ενώ δεν μπορεί να κλιμακώσει, να ανταπεξέλθει σε ποικιλία, έχει χαμηλής ποιότητας προτάσεις, και είναι υπολογιστικά ακριβά [4].

  Μια άλλη προσέγγιση στις συστάσεις είναι μέσω χρήσης βαθειάς μάθησης, όμως είναι δύσκολο να καταλάβουμε πως δουλεύουν αυτά τα μοντέλα και χρειάζονται πάρα πολλά δεδομένα και υπολογισμούς για να κάνουν καλές προβλέψεις.

  Σε αντίθεση με τις κλασικές μεθόδους, η EM μπορεί να διαχειριστεί ακολουθιακές και δυναμικές αλληλεπιδράσεις μεταξύ συστήματος και χρήστη και να λάβει υπ΄όψιν την μακροχρόνια αφοσίωση των χρηστών.

  Η χρήση contextual bandits για

Κεφάλαιο 4

Τι κάναμε εμείς

4.1 Το διαλογικό σύστημα Rasa

4.2 Συστάσεις μέσα στο Rasa

Κεφάλαιο 5

Αποτελέσματα και περαιτέρω εργασία

5.1 Αποτελέσματα

5.2 Επεκτάσεις

Βιβλιογραφία

[1] A. M. Turing, "Computing machinery and intelligence", English, *Mind*, New Series, vol. 59, no. 236, pp. 433–460, 1950, ISSN: 00264423. [Online]. Available: http://www.jstor.org/stable/2251299.

[2] Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας και Η. Σακελλαρίου, *Τεχνητή Νοημοσύνη*. Πανεπιστήμιο Μακεδονίας, 2006.

[3] J. Langford and T. Zhang, "The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information", in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20, Curran Associates, Inc., 2007.

[4] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511763113.

[5] D. Silver, *Lectures on reinforcement learning*, URL: https://www.davidsilver.uk/teaching/, 2015.

[6] P. H. Aditya, I. Budi, and Q. Munajat, "A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for e-commerce in indonesia: A case study pt x", in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 303–308. DOI: 10.1109/ICACSIS.2016.7872755.

[7] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, *Deep reinforcement learning for dialogue generation*, 2016. DOI: 10.48550/ARXIV.1606.01541. [Online]. Available: https://arxiv.org/abs/1606.01541.

[8] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey", *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, Nov. 2017, ISSN: 1558-0792. DOI: 10.1109/MSP.2017.2743240.

[9] N. T. Blog, *Artwork Personalization at Netflix*, en, Dec. 2017. [Online]. Available: https://netflixtechblog.com/artwork-personalization-c589f074ad76 (visited on 01/30/2023).

[10] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on thompson sampling", 2017. DOI: 10.48550/ARXIV.1707.02038. [Online]. Available: https://arxiv.org/abs/1707.02038.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018, ISBN: 0262039249.

[12] Z. Zhang, X. Li, J. Gao, and E. Chen, *Budgeted policy learning for task-oriented dialogue systems*, 2019. DOI: 10.48550/ARXIV.1906.00499. [Online]. Available: https://arxiv.org/abs/1906.00499.

[13] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020, ISBN: 9781108486828.

[14] M. M. Afsar, T. Crump, and B. Far, *Reinforcement learning based recommender systems: A survey*, 2021. DOI: 10.48550/ARXIV.2101.06286. [Online]. Available: https://arxiv.org/abs/2101.06286.

[15] J. Ni, T. Young, V. Pandelea, F. Xue, V. Adiga, and E. Cambria, "Recent advances in deep learning based dialogue systems: A systematic survey", *CoRR*, vol. abs/2105.04387, 2021. arXiv: 2105.04387. [Online]. Available: https://arxiv.org/abs/2105.04387.

[16] L. Ouyang, J. Wu, X. Jiang, *et al.*, *Training language models to follow instructions with human feedback*, 2022. DOI: 10.48550/ARXIV.2203.02155. [Online]. Available: https://arxiv.org/abs/2203.02155.

[17] Z. Yan, "Bandits for recommender systems", *eugeneyan.com*, May 2022. [Online]. Available: https://eugeneyan.com/writing/bandits/.