Εισαγωγή

Στην πρώτη άσκηση προγραμματισμού, υλοποιήθηκαν ο αλγόριθμος ε-Greedy και ο αλγόριθμος Upper Confidence Bound (Πολυπλοκότητας), για την επίλυση ενός προβλήματος multi-arm bandit. Ο στόχος αυτής της άσκησης ήταν να διερευνήσουμε την απόδοση αυτών των αλγορίθμων όσο αφορά το regret, το οποίο είναι η διαφορά μεταξύ της αναμενόμενης ανταμοιβής του καλύτερου χεριού και της αναμενόμενης ανταμοιβής του επιλεγμένου χεριού σε κάθε χρονικό βήμα. Το περιβάλλον που χρησιμοποιήθηκε για το πρόβλημα αυτό αποτελούνταν από k στοχαστικά bandits, όπου έκαστο είχε μια ανταμοιβή που ήταν ομοιόμορφα κατανεμημένη εντός δύο ορίων. Τα όρια αυτά ήταν δύο αριθμοί μεταξύ του 0 και του 1 και ήταν διαφορετικοί για κάθε bandit. Δημιουργήθηκαν διαγράμματα για να αναλυθούν τα ποσοστά regret κάθε αλγορίθμου και έγινε σύγκριση της ταχύτητας μάθησης για T=1000 και k=10, καθώς και για άλλους συνδυασμούς T και k, έτσι ώστε να γίνει κατανοητό το πως επηρεάζει η κάθε παράμετρος την τελική ανταμοιβή και το regret.

Αλγόριθμοι

ε-Greedy

Στο περιβάλλον που δημιουργήθηκε για την διεξαγωγή του πειράματος χρησιμοποιήθηκε ε-Greedy με υπογραμμικό ε, ο αλγόριθμος αυτός είναι ένας αλγόριθμος ενισχυτικής μάθησης που εξισορροπεί την εξερεύνηση και την εκμετάλλευση κατά την επιλογή δράσης επιλέγοντας τη δράση με την υψηλότερη ανταμοιβή με πιθανότητα (1-ε), όπου ε είναι μια μικρή θετική τιμή που μειώνεται με την πάροδο του χρόνου ακολουθώντας ένα υπογραμμικό πρόγραμμα. Η υπογραμμική μείωση του ε διασφαλίζει τη συνεχή εξερεύνηση και τη δυνητική ανακάλυψη νέων ενεργειών υψηλής ανταμοιβής ακόμη και μετά από πολλές επαναλήψεις. Συγκεκριμένα το ε δινόταν από τον τύπο $e_t = O(t^{-\frac{1}{3}} \cdot (klogt)^{\frac{1}{3}})$, όπου k ο αριθμός στοχαστικών bandits και t ο αριθμός του γύρου του υπολογισμού του ε. Με την χρήση αυτού του αλγορίθμου επιτυγχάνεται $O(t^{\frac{2}{3}}(klogt)^{\frac{1}{3}})$ πολυπλοκότητα regret $t \ \forall \ T$.

Upper Confidence Bound

Ο UCB είναι ένας αλγόριθμος στην ενισχυτική μάθηση που επιλέγει ενέργειες με βάση την εκτιμώμενη αξία τους, λαμβάνοντας υπόψη τόσο τη μέση ανταμοιβή όσο και τον βαθμό αβεβαιότητας. Αποδίδει ένα "όριο εμπιστοσύνης" σε κάθε ενέργεια και επιλέγει την ενέργεια με το υψηλότερο ανώτερο όριο εμπιστοσύνης για να εξισορροπήσει την εξερεύνηση και την εκμετάλλευση. Η regret πολυπλοκότητά του είναι $O(\sqrt{ktlogT})$, καθιστώντας το αποτελεσματικό εργαλείο για την εκμάθηση βέλτιστων ενεργειών σε πολύπλοκα περιβάλλοντα με υψηλή αβεβαιότητα. Στην συγκεκριμένη άσκηση κάθε όριο εμπιστοσύνης υπολογίζεται μέσω του τύπου

Regret

Σε αυτό το κομμάτι της αναφοράς συγκρίνονται οι θεωρητικές πολυπλοκότητες regret καθεμιάς από τους δύο αλγορίθμους με τις πειραματικές που υπολογίζονται από τον κώδικα.

Figure 1

Παρατηρείται ότι καθώς αυξάνεται ο αριθμός των γύρων, η κλίση των καμπύλης που προκύπτει για τον κάθε αλγόριθμο μειώνεται, όχι με τον ίδιο βαθμό. Αυτό υποδηλώνει ότι η υπολογιστική πολυπλοκότητα αυτών των αλγορίθμων μπορεί να περιγραφεί κατάλληλα ως υπογραμμική.

Αποτελέσματα

Πραγματοποιήθηκαν παρατηρήσεις σχετικά με την απόδοση κάθε αλγορίθμου μέσω της συλλογής δεδομένων για διαφορετικές τιμές των ανεξάρτητων μεταβλητών T και k. Συγκεκριμένα, ελήφθησαν τρία σύνολα δεδομένων, δηλαδή T = 1000 και k = 10, T = 10000 και k = 10 και T = 1000 και k = 5. Οι τιμές αυτές επιλέχθηκαν με στόχο τη διερεύνηση της επίδρασης των ανεξάρτητων μεταβλητών στην απόδοση του αλγορίθμου. Η πρώτη περίπτωση χρησιμοποιείται ως σημείο αναφοράς και μετά παρατηρείται η αλλαγή του αποτελέσματος με την αλλαγή των μεταβλητών. Ακολουθούν τα αποτελέσματα από διάφορων παραλλαγών T και k.

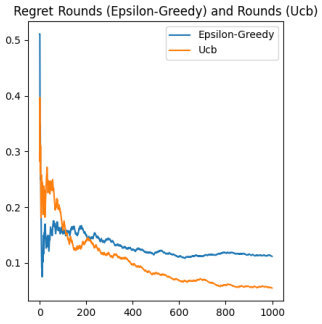

Figure 2: Bandit=5 and rounds=1000

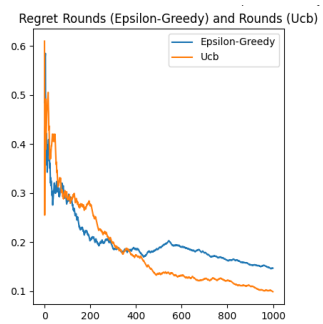

Figure 3: Bandit=10 and rounds=1000

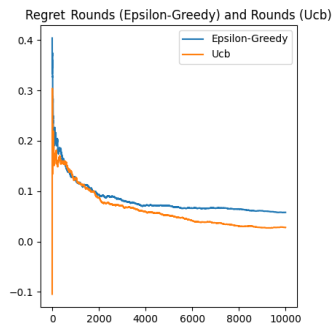

Figure 4: Bandit=10 and rounds=10000

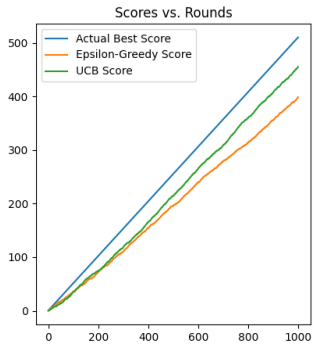

Figure 5: Bandit=5 and rounds=1000

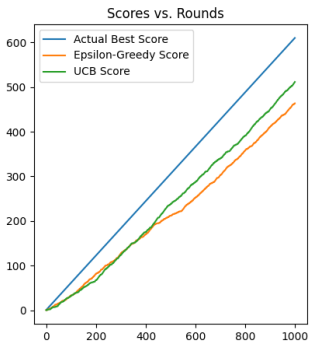

Figure 6: Bandit=10 and rounds=1000

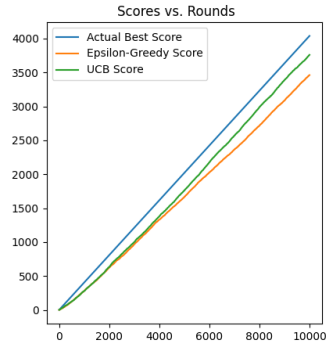

Figure 7: Bandit=10 and rounds=10000

Πραγματοποιώντας μια συγκριτική ανάλυση των μεταβλητών προφοράς, συγκεκριμένα με k = 10 και T = 1000, είναι σαφές ότι ο αλγόριθμος UCB υπερέχει των αντίστοιχων αλγορίθμων του στην προσέγγιση των βέλτιστων αποτελεσμάτων. Καθώς κάθε αλγόριθμος προχωρά σε επόμενους γύρους, η τιμή regret μειώνεται, αλλά ο αλγόριθμος UCB επιδεικνύει μεγαλύτερη μείωση. Η επίδραση κάθε μεταβλητής στη συμπεριφορά του αλγορίθμου αποτελεί μια ενδιαφέρουσα παρατήρηση. Είναι προφανές ότι καθώς αυξάνεται ο αριθμός των γύρων, και οι δύο αλγόριθμοι εμφανίζουν μειωμένο regret και συγκλίνουν προς το βέλτιστο αποτέλεσμα, ενώ κάτι αντίστοιχο μπορεί να ειπωθεί και για μικρότερο αριθμό χεριών. Αυτό μπορεί να παρατηρηθεί και στα διαγράμματα Figure 5,6 και 7, καθώς στο figure 5 αι 7 οι γραμμές score των αλγορίθμων βρίσκονται πιο κοντά στο επιθυμητό αποτέλεσμα. Αναμενόμενο να υπάρχει ανάλογη σχέση μεταξύ πλήθος γύρων και σύγκλισης στο βέλτιστο score, καθώς περισσότερες επαναλήψεις επιτρέπουν στους αλγορίθμους να εντοπίζουν τις καλύτερα bandits, με αποτέλεσμα να ελαχιστοποιείται η τιμή του regret. Ομοίως και με το αριθμό χεριών, όσο λιγότερα υπάρχουν τόσο αυξάνεται η πιθανότητα επιλογής του χεριού με την καλύτερη μέση αντα-μοιβή. Υπάρχει μια μικρή πιθανότητα στην περίπτωση με τα 5 bandits να βρεθεί ο epsilon-Greedy με καλύτερο αποτέλεσμα από το Ucb, πράγμα το οποίο είναι προϊόν καθαρής τύχης καθώς ο αλγόριθμος Ucb κατά κανόνα είναι πιο αποδοτικός.