Reinforcement Learning

ε-Greedy and UCB

Σταυρόπουλος Αλέξανδρος Ανδρέας
2019030109

Διδάσκων:
Θρασύβουλος Σπυρόπουλος

Υπεύθυνος εργαστηρίου:
-

ΗΜΜΥ
Πολυτεχνείο Κρήτης
Χειμερινό εξάμηνο 2022-2023

Εισαγωγή

Στην πρώτη άσκηση ζητείται η υλοποίηση των αλγορίθμων ε-Greedy και Upper Confidence Bound, οι οποίοι επιτυγχάνουν ισορροπία μεταξύ exploration και exploitation στο γνωστό πρόβλημα κουλοχέρηδων (Bandits Problem). Εφόσον υλοποιήθηκε κάθε αλγόριθμος, δόθηκε ως όρισμα το ίδιο σύνολο από bandits ώστε να εκτελεστούν για συγκεκριμένο αριθμό γύρων (Ορίζοντας T) και επίσης επιλέχθηκε ένας από τους bandits με βάση τον γινόμενο μεταξύ reward και πιθανότητας επιτυχίας. Η επίδοση κάθε αλγορίθμου βασίζεται στο regret το οποίο είναι η διαφορά μεταξύ του σκορ του αρχικά επιλεγμένου bandit και του σκορ που μάζεψε ο κάθε αλγόριθμος.

ε-Greedy

Στον αλγόριθμο ε-Greedy στην περίπτωση exploration επιλέγεται τυχαίο χέρι ανεξάρτητα των επιδόσεών του έως τώρα ενώ στην περίπτωση exploitation επιλέγεται το χέρι με την καλύτερη επίδοση. Η επίδοση κάθε χεριού ορίζεται μέσω του συντελεστή μ ο οποίος είναι ίσος με τον πηλίκο μεταξύ του σκορ που έχει μαζέψει το εκάστοτε χέρι προς τον αριθμό των φορών που έχει επιλεχθεί. Το σκορ προκύπτει ως το γινόμενο μεταξύ reward και διωνυμικής πιθανότητας κάθε χεριού και υπολογίζεται σε κάθε γύρο ξεχωριστά.

Η επιλογή μεταξύ exploitation και exploitation γίνεται με χρήση της μεταβλητής $epsilon = (t)^{-\frac{1}{3}} \cdot (k \cdot \log(t))^{\frac{1}{3}}$. Σε κάθε γύρο, με πιθανότητα epsilon γίνεται explore ενώ με πιθανότητα 1 - epsilon γίνεται exploit και λαμβάνοντας υπόψιν τον φθίνον ρυθμό της μεταβλητής epsilon, μπορεί εύκολα να αποδειχθεί πως όσο αυξάνονται οι γύροι τόσο πιο πιθανό είναι να γίνει exploitation χρησιμοποιώντας το καλύτερο χέρι ενώ ταυτόχρονα τόσο πιο απίθανο να επιλεγεί κάποιο τυχαίο (exploration). Η υλοποίηση παρουσιάζει convergence rate ίσο με $O\left((t)^{\frac{2}{3}} \cdot (K \cdot \log(t))^{\frac{1}{3}}\right)$

Upper Confidence Bound

Στον αλγόριθμο Upper Confidence Bound (UCB) η επιλογή χεριού γίνεται με βάση τον συντελεστή $ucb = \mu + \sqrt{\frac{\log T}{Q}}$ όπου μ ο συντελεστής επίδοσης κάθε χεριού, T ο ορίζοντας και Q οι φορές που έχει επιλεχθεί το αντίστοιχο χέρι. Σε κάθε γύρο επιλέγεται το χέρι το οποίο παρουσιάζει μεγαλύτερο συντελεστή ucb κάτι το οποίο προκύπτει είτε λόγο καλής επίδοσης (συντελεστής μ) είτε επειδή έχουν περάσει πολλοί γύροι που δεν έχει επιλεχθεί το εκάστοτε χέρι. Αυτό έχει ως αποτέλεσμα, το exploration πρακτικά να μην σταματάει ποτέ καθώς ακόμα και για μεγάλο αριθμό γύρου, ανά διαστήματα επιλέγεται διαφορετικό χέρι και αξιολογείται εκ νέου το performance του κάτι το οποίο στον ε-Greedy είναι σχεδόν απίθανο να συμβεί. Η υλοποίηση παρουσιάζει convergence rate ίσο με $O\left(\sqrt{K \cdot T \cdot \log T}\right)$

Σύγκριση επιδόσεων

Εξετάζοντας τα convergence rate κάθε αλγορίθμου αναμένεται η επίδοση του UCB να είναι καλύτερη σε σχέση με του ε-Greedy. Για την επαλήθευση αυτής της εκτίμησης, έγινε σύγκριση των επιδόσεων μεταξύ των δύο αλγορίθμων για το ίδιο σετ bandits σε κάθε τεστ. Αρχικά, για 10 bandits και ορίζοντα μεγέθους 1000, τα αποτελέσματα που εξήχθησαν δεν συναδουν πάντα με την εκτίμηση. Συχνότερη περίπτωση αποτελεί, όπως ήταν αναμενόμενο, ο UCB να παρουσιαζει καλύτερο performance δηλαδή μικρότερο regret (fig:1), ωστόσο, υπήρξαν σπάνιες περιπτώσεις όπου ο ε-Greedy παρουσίασε καλύτερη επίδοση (fig:2) κάτι το οποίο είναι πιθανό να συμβεί καθώς το σκορ κάθε αλγορίθμου υπολογίζεται ξεχωριστά οπότε για τυχαίους λόγους ο ένας αλγόριθμος να είναι πολύ πιο "τυχερός" από τον άλλο.

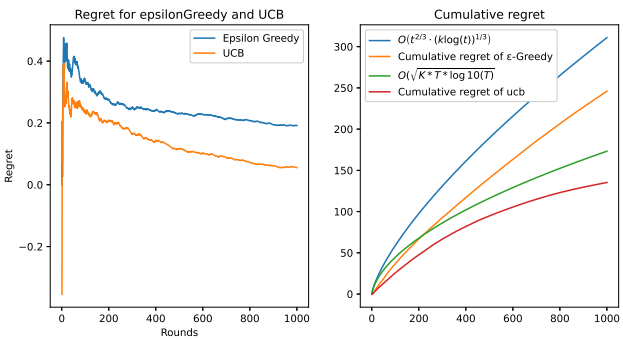

Figure 1: Lower regret for UCB (K=10 T=1000)

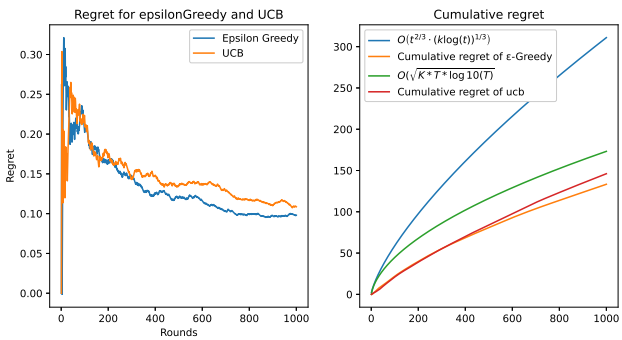

Figure 2: Lower regret for ε-Greedy (K=10 T=1000)

Παρατηρώντας και το cumulative regret, στην σύγκριση μεταξύ των 2 αλγορίθμων οι τιμές του κάθε regret είναι οι αναμενόμενες για την εκάστοτε περίπτωση, όπως και στην σύγκριση κάθε αλγόριθμου με το θεωρητικό του convergence rate. Το θεωρητικό convergence rate αποτελεί πάνω όριο εφόσον εμπεριέχονται τυχαίες μεταβλητές και όπως ήταν αναμενόμενο η γραφική του είναι πιο πάνω σε σχέση με τα πειραματικά δεδομένα.

Για την καλύτερη επαλήθευση της εκτίμησης, αυξήθηκε η τιμή ορίζοντα (T = 10000) ενώ ο αριθμός των bandit παρέμεινε σταθερός (K = 10). Στην περίπτωση αυτή ο UCB πάντα παρουσιάζει καλύτερη επίδοση, δηλαδή μικρότερο regret σε σχέση με τον ε-Greedy, κάτι το οποίο επιβεβαιώνει την αρχική εκτίμηση.

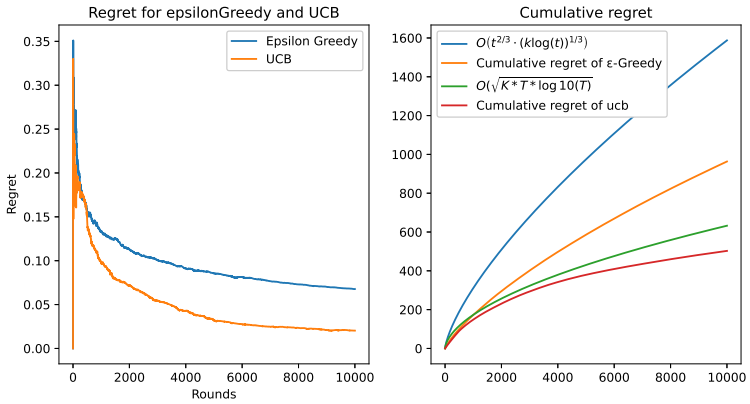

Figure 3: Lower regret for ε-Greedy (K=10 T=10000)

Παρατηρώντας συνολικά τις γραφικές, είναι εμφανές πως όσο αυξάνεται ο αριθμός των γύρων τόσο μειώνεται το regret αλλά και πως όσο αυξάνονται οι γύροι ο ρυθμός αύξησης του του regret μειώνεται και για του δύο αλγορίθμους κάτι το οποίο σημαίνει πως και οι δύο αλγόριθμοι επιτυγχάνουν sublinear πολυπλοκότητα.

Τέλος, έγινε μεταβολή του αριθμού των bandits και πιο συγκεκριμένα τέθηκαν ίσοι με 5 και 20, για ορίζοντα 1000 και 10000 αντίστοιχα.

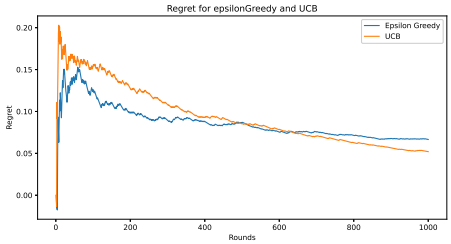

Figure 4: K = 5, T = 1000

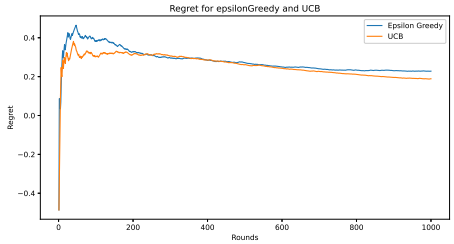

Figure 5: K = 20, T = 1000

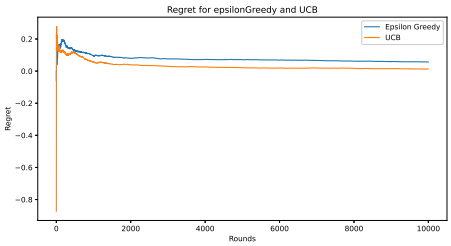

Figure 6: K = 5, T = 10000

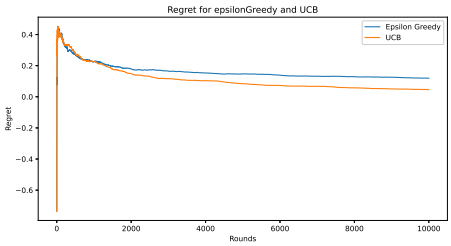

Figure 7: K = 20, T = 10000

Όπως ήταν αναμενόμενο στις περιπτώσεις που μειώνεται ο αριθμός των bandits, το regret σταθεροποιείται σε μικρότερη τιμή, ενώ αντίθετα αυξάνοντας τον αριθμό των bandits, αυξάνεται η τιμή σταθεροποίησης.