Programming Assignment 1

Ζαφειράκης Κωνσταντίνος 2019030035

Εισαγωγή

Μας ζητήθηκε να εφαρμόσουμε για Multi Armed Bandits με τυχαίο reward τους αλγορίθμους ε-Greedy και UCB. Σε κάθε χρονικό βήμα, ο αλγόριθμος ε-Greedy επιλέγει ένα τυχαίο arm με πιθανότητα ε και το πίο κερδοφόρο arm με πιθανότητα (1-ε). Αυτή η ισορροπία μεταξύ εξερεύνησης και εκμετάλλευσης επιτρέπει στον αλγόριθμο να συλλέγει περισσότερες πληροφορίες για το περιβάλλον, μεγιστοποιώντας παράλληλα την αναμενόμενη ανταμοιβή. Η τιμή του ε καθορίζει το συμβιβασμό μεταξύ εξερεύνησης και εκμετάλλευσης. Το ε που εφαρμόζουμε εμείς ισούται με $t^{-\frac{1}{3}} \cdot (k \cdot log\,(t))^{\frac{1}{3}}$ είναι μεταβλητό και όλο μικραίνει ώστε να έχουμε περισσότερες πιθανότητες εκμετάλευσης όσο περνάει ο χρόνος και ξέρουμε περισσότερο για το κάθε bandit, η πολυπλοκότητα για αυτό το ε ισούται με $O\left(t^{\frac{2}{3}} \cdot (k \cdot log\,(t))^{\frac{1}{3}}\right)$. Ο UCB σε κάθε χρονικό βήμα, επιλέγει το arm με το υψηλότερο άνω όριο εμπιστοσύνης, το οποίο εξισορροπεί την αναμενόμενη ανταμοιβή της ενέργειας με την αβεβαιότητα της εκτιμώμενης τιμής. Κάθε βήμα προσαμρόζει το άνω όριο εμπιστοσύνης ανάλογα με δεδομένα που μαζέυει απο τα arms. Τέλος έχει πολυπλοκότητα $O(\sqrt{k \cdot t \cdot log\,(t)})$ είναι δηλαδή θεωρητικά πίο αποτελεσματικός απο τον ε-Greedy.

Θεωρητική-Πειραματική Συγκριση

Θεωρητικά οι αλγόριθμοι ε-Greedy και UCB έχουν πολυπλοκότητες $O\left(t^{\frac{2}{3}} \cdot (k \cdot log\,(t))^{\frac{1}{3}}\right)$ και $O(\sqrt{k \cdot t \cdot log\,(t)})$ αντιστοίχως. Πειραματικά τις περισσότερες φορές τα αποτελέσματα συμβαδίζουν με την θεωρία καθώς το αθροιστικό regret όντως δέν ξεπερνάει το άνω όριο της πολυπλοκότητας όπως φαίνεται και στα figure 1 και figure 2. Εφόσον όμως το πρόβλημα έχει τυχαίες μεταβλητές υπάρχουν περιπτώσεις όπου το αθροιστικό regret υπερβαίνει το άνω όριο αυτό μπορεί να συμβέι και στους δύο αλγορίθμους ακόμη όσο μικρότερο το horizon τόσο πιο πιθανό όπως φαίνεται και στο figure 3. Ανεξαρτήτως όμως οι αλγόριθμοι έιναι sublinear αφου όσο περνάει χρόνος μαθαίνουν κάτω που φαίνετε στα παρακάτω figures δηλαδή όσο προχωράει ο χρόνος το αθροιστικό regret μειώνει σημαντικά την κλίση του

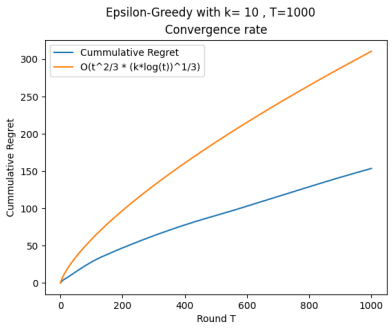

Figure 1: ε-Greedy convergence

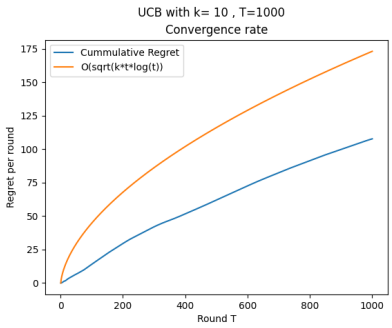

Figure 2: UCB convergence

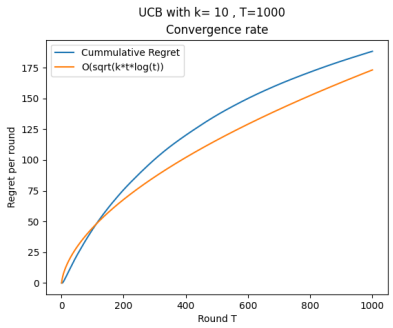

Figure 3: UCB convergence

Σύγκριση αλγορίθμων

Θεωρητικά περιμένουμε ο αλγόρθμος UCB να είναι πιο αποδοτικός από τον ε-Greedy όπως και ισχύει για k=10 και T=1000 καθώς το regret του UCB είναι μικρότερο από αυτό του ε-Greedy η γραφική παράσταση φαίνεται και παρακάτω στο figure4. Στην συνέχεια όπως ζητείται πραγματοποιούνται 2 ακόμη τεστ το ένα με αυξημένο αριθμό χεριών (k=100,T=1000) όπως φαίνεται στο figure 5 και το άλλο με αυξημένο ορίζοντα T (k=10,T=100000) όπως φαίνεται και στο figure 6.

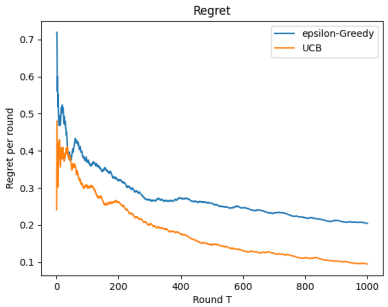

Figure 4: k=10,T=1000

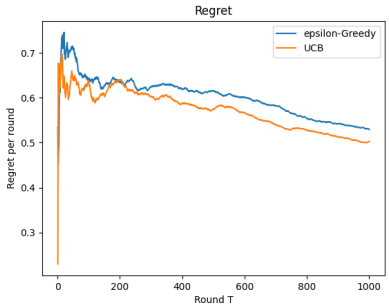

Figure 5: k=100,T=1000

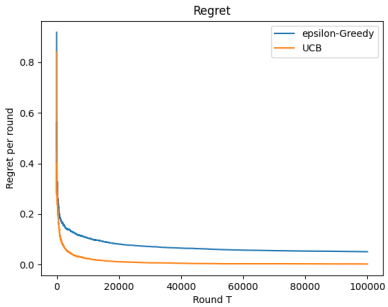

Figure 6: k=10,T=100000

Παρατηρείται αρχικά για k=10,T=1000 ότι στην αρχή τα regret των UCB με τον ε-Greedy δεν απέχουν πολύ άλλα όσο αυξάνονται τα round τόσο αυξάνονται. Στην περίπτωση k=100,T=1000 παρατηρούμε ότι το regret και των δύο αλγορίθμων έχει αυξηθεί σημαντικά αυτό συμβαίνει γιατί με την προσθήκη χεριων η δυσκολία εύρεσης του αποδοτικότερου χεριού αυξάνεται συμαντικά καθώς πρέπει όλα τα χέρια να εξεταστούν ως ενα βαθμό. Ακόμη παρατηρείται ότι ποσοστιαία τα regret του UCB και ε-Greedy είναι αρκετά μικρότερα σε σχέση με αυτά της πρώτης περίπτωσης αυτό συμβαίνει γιατί ο ορίζοντας 1000 rounds δέν είναι αρκετός για τους αλγόριθμους ειδικά για τον UCB να κάνουν exploit σε μεγαλύτερο βαθμό γιατί ακόμη ερευνούν σε σημαντικό ρυθμό τα χέρια εάν είχαμε πχ k=100/T=10000000 τότε θα βλέπαμε τον κάθε αλγόριθμο να λειτουργεί κοντά στο βέλτιστο του και να υπάρχει μεγαλύτερη διαφορά μεταξύ UCB και ε-Greedy στα τελευταία rounds. Παρόλα αυτά ο UCB πάλι έχει καλύτερη απόδοση απο τον ε-Greedy. Τέλος για k=10/T=100000 παρατηρούμε ότι τα τελικά regrets είναι ελάχιστα αυτό του ε-Greedy είναι κοντά στο 0.5 ενώ αυτό του UCB γύρω στο 0.03 και εκτός αυτού φαίνεται να έχουν σταθεροποιηθεί σε αυτές τις τιμές αυτό συμβαίνει γιατι και οι δύο αλγόριθμοι μετά απο τόσα rounds έχουν βρεί τα καλύτερα χέρια και παίζουν αυτά, δουλεύουν δηλαδή κοντά στην βέλτιστη απόδοση τους κάτι το οποίο δεν το είδαμε στα δύο προηγούμενα graphs καθώς δεν είχαν σταθεροποιηθεί προς το τέλος αλλά συνέχιζαν να πέφτουν. Τέλος πάλι ο UCB είναι αποδοτηκότερος σε αυτήν την περίπτωση. Αξίζει να σημειωθεί οτι καθώς οι αλγόριθμοι εφαρμόζονται σε σύστημα με τυχαίες τιμές δέν αποκλείεται κυρίως σε μικρό horizon να υπάρχουν περιπτώσεις που ο ε-Greedy βγάζει καλύτερα αποτελέσματα απο τον UCB καθαρά λόγω τύχης είναι πιθανόν λοιπόν σε ενα μικρό αριθμό graphs με ορίζοντα 1000 να υστερεί ο UCB σε μεγάλο ορίζοντα όμως όπως στην τελευταία περίπτωση η πιθανότητα να υστερεί ο UCB είναι αφάνταστα μικρή καθώς υπάρχουν πολλά rounds στα οποία ελέγχονται τα χέρια.