Αλγοριθμικές Τεχνικές για Δεδομένα Ευρείας Κλίμακας

1ο Φυλλάδιο Ασκήσεων

Προθεσμία υποβολής: 10/04/2025, 23:59 (ώρα Ελλάδος)

Ενότητα Α: Μετρητές Morris

Θεωρούμε έναν μετρητή τύπου Morris που εκτιμά το πλήθος στοιχείων n σε μια ροή δεδομένων (εκτιμώντας το πλήθος των bits στην δυαδική αναπαράσταση του n), ο οποίος περιγράφεται από τον εξής ψευδοκώδικα:

```
integer C ← 0
void insert()
{
   C ← C+1, with probability 1/2^C
}
```

```
integer query()
{
   return 2^C - 1
}
```

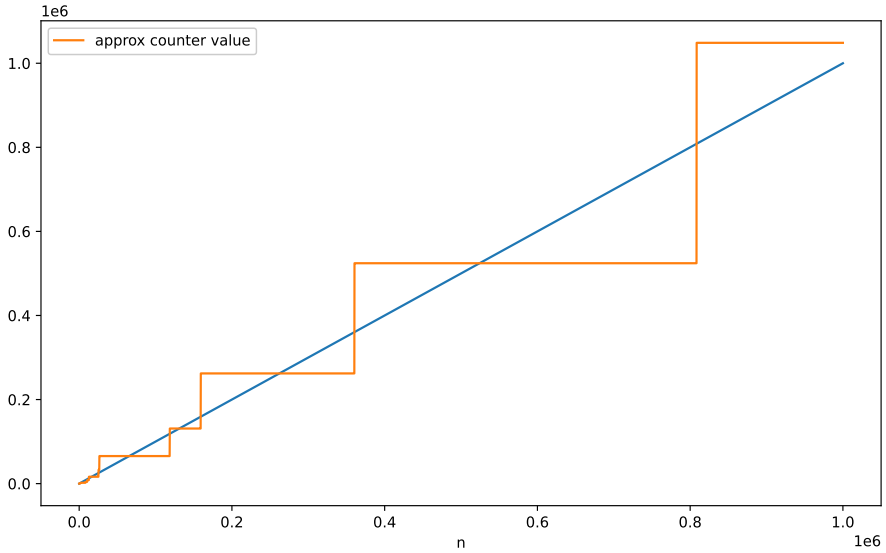
Ορίζουμε το C_n να είναι η τυχαία μεταβλητή που η τιμή της είναι η τιμή της μεταβλητής C μετά από n εισαγωγές στοιχείων.
(Παρατηρήστε ότι $C_n \in \{1, 2, \ldots, n\}$.)

Άσκηση 1

α) *(Μονάδες: 0.75)*
Υλοποιήστε τον παραπάνω μετρητή Morris, και τρέξτε τον για 1.000.000 εισαγωγές στοιχείων. Έπειτα από κάθε εισαγωγή, εκτυπώστε την εκτίμηση του μετρητή (δηλαδή το $2^C - 1$).

Σχεδιάστε μία αντίστοιχη γραφική παράσταση που δείχνει τις τιμές του μετρητή καθώς γινόντουσαν οι εισαγωγές. Το αποτέλεσμά σας θα πρέπει να είναι όπως φαίνεται στην ακόλουθη εικόνα*:

*προφανώς η γραφική παράσταση θα ποικίλλει ανάλογα με την αρχικοποίηση που κάνατε στην συνάρτηση που παράγει την τυχαιότητα



Παράδειγμα εκτιμήσεων του μετρητή Morris για το πλήθος εισαγωγών n, όσο το n αυξάνεται από 1 μέχρι και 1.000.000.

β1) *(Μονάδες: 0.75)*
Μπορούμε να βελτιώσουμε τον παραπάνω μετρητή ως εξής. Αντί να κρατούμε μόνο μία μεταβλητή C, μπορούμε να κρατούμε αρκετές τέτοιες μεταβλητές (οι οποίες ανανεώνονται ανεξάρτητα), και να επιστρέφουμε τον μέσο όρο ή τον διάμεσο από το αντίστοιχο αποτέλεσμα που δίνουν. (Π.χ., αν κρατούμε τις μεταβλητές C1, C2, C3, τότε επιστρέφουμε ως εκτίμηση του n το $(2^{C1} + 2^{C2} + 2^{C3} - 3)/3$.)

Υλοποιήστε αυτήν την ιδέα, για τον μέσο όρο και για τον διάμεσο, χρησιμοποιώντας 5 ανεξάρτητες μεταβλητές, και κάντε τις αντίστοιχες γραφικές παραστάσεις όπως και στο α).

β2) *(Μονάδες: 0.25)*
Δοκιμάστε να τρέξετε αρκετές φορές τις υλοποιήσεις που κάνατε στο β1 (χρησιμοποιώντας κάθε φορά νέα τυχαιότητα). Θεωρείτε ότι είναι καλύτερο να επιστρέφετε τον μέσο όρο ή τον διάμεσο; (Ίσως χρειαστεί να προσθέσετε περισσότερες μεταβλητές C για να παρατηρήσετε ένα πιο ξεκάθαρο μοτίβο.)

Προφανώς είναι ανούσιο στην πράξη να κρατά κανείς πέντε διαφορετικές μεταβλητές C για να εκτιμήσει το πλήθος στοιχείων όταν το n φτάνει μέχρι το 1.000.000. Πρώτον, τεκμηριώστε αυτόν τον ισχυρισμό. Δεύτερον, για ποια n θα είχε νόημα να κρατά κανείς 5 μεταβλητές C, δεδομένου ότι θέλουμε να έχουμε κέρδος στον χώρο; (Θεωρήστε ότι μόνο οι μεταβλητές C είναι αυτές που κοστίζουν χώρο μνήμης.)

(Υπόδειξη: χρειαζόμαστε $\lceil \log_2(n+1) \rceil$ bits για να αναπαραστήσουμε αριθμούς από το 0 μέχρι και το n.)

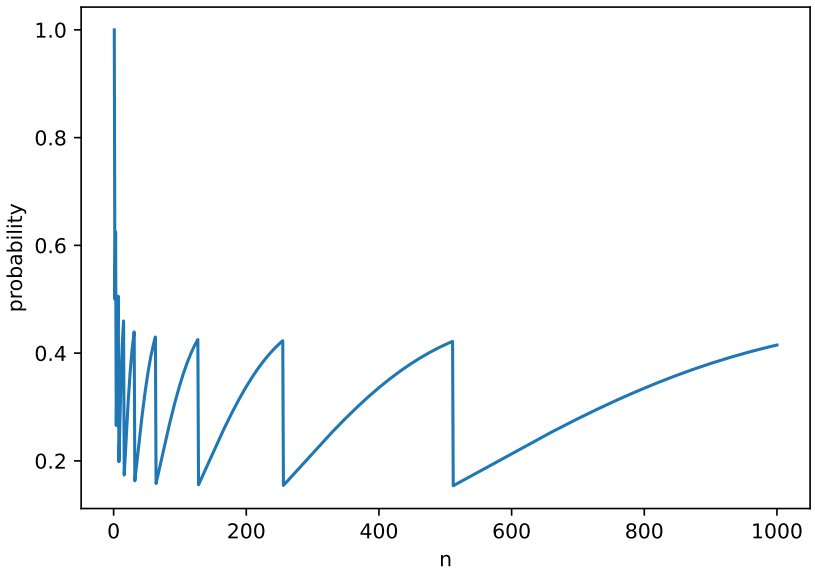
Άσκηση 2

α) *(Μονάδες: 0.125)*

Υπολογίστε με απόλυτη ακρίβεια την πιθανότητα η C_{1000} να έχει τιμές στο σύνολο {8, 9, 10, 11}. Δηλαδή, αυτή η πιθανότητα θα πρέπει να υπολογιστεί ως κλάσμα ακεραίων. Να αναφέρετε το αποτέλεσμα σε δεκαδική μορφή (με ακρίβεια τουλάχιστον 5 δεκαδικών ψηφίων), και υποβάλετε τον κώδικα με τον οποίο υπολογίσατε αυτήν την πιθανότητα.

β) *(Μονάδες: 0.125)*

Υπολογίστε τις πιθανότητες $P(C_k = \lceil \log_2(k+1) \rceil)$, για $k \in \{1, \ldots, 1000\}$. Σχεδιάστε μια γραφική παράσταση που αναπαριστά αυτές τις πιθανότητες. Το αποτελεσμά σας θα πρέπει να είναι κάπως έτσι:



Θα πρέπει να υποβάλετε τον κώδικα με τον οποίο υπολογίσατε τις πιθανότητες, και τον κώδικα με τον οποίο φτιάξατε την γραφική παράσταση.

γ) *(Μονάδες: 0.125)*

Κάντε ακριβώς ό,τι και στο β), μόνο που αυτήν την φορά θα πρέπει να υπολογίσετε τις πιθανότητες $P(C_k \in \{l(k)-1, l(k), l(k)+1\})$, όπου $l(k) = \lceil \log_2(k+1) \rceil$, για $k \in \{1, 2, \ldots, 1000\}$.

Αναφέρετε κάποιο χρήσιμο συμπέρασμα που βγαίνει από αυτήν την γραφική παράσταση, ειδικά αν την συγκρίνουμε με αυτήν που λάβαμε στο β).

δ) *(Μονάδες: 0.625)*
Μέχρι τώρα υπολογίσαμε ορισμένες πιθανότητες της μορφής $P(C_k = i)$. Αυτό μας λέει την πιθανότητα, αν γίνουν k εισαγωγές στοιχείων, η μεταβλητή C_k (που θέλουμε να είναι κοντά στο $\lceil \log_2(k+1) \rceil$) να έχει την τιμή i. Αυτή η πιθανότητα όμως δεν μας λέει ακριβώς τι γινόταν *καθόλη την πορεία των εισαγωγών*. Δηλαδή, μπορεί μεν η C_k να τύχει να είναι κοντά στο $\lceil \log_2(k+1) \rceil$, όμως μπορεί για ορισμένα μικρότερα k' η C να ήταν πολύ μακριά απ' το $\lceil \log_2(k'+1) \rceil$.

Εργαστείτε όπως και στα β) και γ), μόνο που αυτήν την φορά υπολογίστε την πιθανότητα:
1) η C_k να είναι ακριβώς $l(k)$ *για όλα τα* $k \in \{1, 2, \ldots, 1000\}$, και
2) η C_k να ανήκει στο σύνολο $\{l(k)-1, l(k), l(k)+1\}$, *για όλα τα* $k \in \{1, 2, \ldots, 1000\}$,
όπου $l(k) = \lceil \log_2(k+1) \rceil$.

Άσκηση 3 *(Μονάδες: 0.75)*

Έστω ότι θέλουμε έναν μετρητή τύπου Morris για να μετρήσουμε μέχρι το 1.000.000, και έστω ότι έχουμε 8 bits μνήμης στην διάθεσή μας. Παρατηρήστε ότι αυτά είναι περισσότερα απ' όσα χρειαζόμαστε (και τεκμηριώστε αυτόν τον ισχυρισμό). Θα θέλαμε λοιπόν να εκμεταλλευτούμε όλα τα διαθέσιμα bits.

Για τον σκοπό αυτό, μπορούμε να κάνουμε την εξής τροποποίηση στον αλγόριθμο. Αντί να αυξάνουμε το C με πιθανότητα $1/2^C$, μπορούμε να το αυξάνουμε με πιθανότητα $1/\alpha^C$, για κάποιο α με $1 \leq \alpha \leq 2$. (Παρατηρήστε ότι για $\alpha = 1$ έχουμε απλώς κανονική καταμέτρηση, με απόλυτη ακρίβεια.) Έπειτα, μια εύλογη επιλογή είναι να επιστρέψουμε ως εκτίμηση του n το $(1/(\alpha - 1)) \cdot (\alpha^C - 1)$ (γιατί;).

Βρείτε ένα κατάλληλο α ώστε να αξιοποιηθούν όσο το δυνατόν περισσότερο τα 8 bits, φτιάξτε μια γραφική παράσταση όπως και στην Άσκηση 1, και εξετάστε αν όντως υπάρχει κάποια βελτίωση στην ποιότητα της προσέγγισης.