Άσκηση 1

1 Εκτίμηση Κατάταξης

Ας υποθέσουμε ότι έχουμε μια αταξινόμητη λίστα διακριτών αριθμών $x_1, x_2, \ldots, x_n$. Θέλουμε να επιλέξουμε τυχαία $m < n$ στοιχεία από αυτήν τη λίστα (με αντικατάσταση) και με βάση αυτό το δείγμα θέλουμε να επιστρέψουμε κάποιο στοιχείο x έτσι ώστε το $rank(x)$ είναι περίπου ίσο με k για ένα δεδομένο k. Με τον όρο $rank$ εννοούμε την κατάταξη του x στην αρχική λίστα. Το μεγαλύτερο από τα $x_1, x_2, \ldots, x_n$ έχει $rank$ 1, η δεύτερη μεγαλύτερη έχει $rank$ 2 κ.λ.π. και προσπαθούμε να βρούμε κάτι με $rank(x) \approx k$. Περιγράψτε μια απλή στρατηγική για την επιλογή ενός τέτοιου x.

Πόσο μεγάλο πρέπει να θέσουμε το m (με συμβολισμό big-O) έτσι ώστε, με πιθανότητα $(1-\delta)$, η στρατηγική σας να επιστρέφει ένα x με $(1-\varepsilon)k \leq rank(x) \leq (1+\varepsilon)k$;

2 Καταμέτρηση Μέσω Ερωτημάτων

Ο φίλος σας έχει στο μυαλό του ένα σύνολο αριθμών $S \subseteq \{1, \ldots, n\}$. Εάν τον ρωτήσετε για ένα σύνολο $Q \subseteq \{1, ..., n\}$ θα απαντήσει αν η τομή $S \cap Q$ είναι άδεια ή όχι. Ο στόχος σας είναι να υπολογίσετε (περίπου) πόσα στοιχεία περιέχονται στο σύνολο S.

(α) Σχεδιάστε μια στρατηγική που να διακρίνει εάν το S περιέχει $\leq k$ ή $\geq (1+\varepsilon)k$ για οποιοδήποτε $k \in \{1, \ldots, n\}$. Δείξτε ότι $O\left(\log(1/\delta)/\varepsilon^2\right)$ ερωτήματα επαρκούν για να διακριθούν οι 2 περιπτώσεις με πιθανότητα $1-\delta$.
Συμβουλή: Επιλέξτε ένα τυχαίο σύνολο Q που περιέχει κάθε στοιχείο $i \in \{1, \ldots, n\}$ με πιθανότητα $1/k$.

(β) Δείξτε ότι αυτό συνεπάγεται έναν αποδοτικό εκτιμητή $\hat{s}$ για το μέγεθος $|S|$, που με πιθανότητα τουλάχιστον $2/3$ ικανοποιεί $\hat{s} \leq |S| \leq (1+\varepsilon)\hat{s}$ χρησιμοποιώντας την προαναφερθείσα στρατηγική. Ο αριθμός των ερωτημάτων θα πρέπει να είναι το πολύ λογαριθμικός στο n.

(γ) Υποθέτοντας ότι γνωρίζετε το μέγεθος $|S|$, σχεδιάστε ένα αποδοτικό σχήμα που δειγματοληπτεί ένα ομοιόμορφα τυχαίο στοιχείο από το S ρωτώντας τον φίλο σας πολλές φορές με διαφορετικά σύνολα Q. Ο αναμενόμενος αριθμός ερωτημάτων θα πρέπει να είναι το πολύ λογαριθμικός στο n.
Συμβουλή: Μπορείτε να βρείτε ένα τυχαίο σύνολο Q που περιέχει *ακριβώς ένα* στοιχείο από το S; Μπορείτε να σκεφτείτε μια αποδοτική μέθοδο για να ελέγξετε εάν το Q ικανοποιεί αυτήν την ιδιότητα, και εάν την ικανοποιεί, να επιστρέψετε το μοναδικό στοιχείο του $S \cap Q$;

3 Απόσταση Kolmogorov

Η απόσταση Kolmogorov μεταξύ των συναρτήσεων μάζας πιθανότητας $p, q : [n] \to [0,1]$ ορίζεται ως $d_K(p,q) \triangleq \max_{i\in[n]} |p([1,i]) - q([1,i])|$, όπου $p([1,u]) \triangleq \sum_{k=1}^{i} p(k)$.

(α) Δώστε έναν αλγόριθμο για να μάθετε μια αυθαίρετη κατανομή στο $[n]$ σε Kolmogorov απόσταση ε με πιθανότητα $1-\delta$ χρησιμοποιώντας $O(\log(1/\varepsilon\delta)/\varepsilon^2)$ δείγματα. Πως εξηγείτε το γεγονός ότι ο αλγόριθμός σας έχει δειγματική πολυπλοκότητα ανεξάρτητη από το n;
Extra Credit: Μπορείτε να αφαιρέσετε τον όρο $\log(1/\varepsilon)$ για να πάρετε την βέλτιστη πολυπλοκότητα;

(β) Δείξτε ότι οποιοσδήποτε αλγόριθμος εκμάθησης για το παραπάνω πρόβλημα απαιτεί $\Omega(\log(1/\delta)/\varepsilon^2)$ δείγματα.

(γ) Έστω $p : [n] \to [0,1]$ μια συνάρτηση μάζας πιθανότητας που γνωρίζουμε ότι είναι μονότονη (μη αύξουσα) στο διατεταγμένο πεδίο της, δηλ. $p_{i+1} \leq p_i$ για όλα τα $i \in [n-1]$. Σχεδιάστε έναν αλγόριθμο που διακρίνει με πιθανότητα τουλάχιστον $1-\delta$ τις περιπτώσεις που η p είναι ομοιόμορφη, δλδ $p = U_n$, και $d_{TV}(p, U_n) \geq \varepsilon$. Δείξτε ότι ο αλγόριθμός σας είναι βέλτιστος ως προς τη δειγματική του πολυπλοκότητα.