Υπογραμμικοί Αλγόριθμοι
Μάθημα 6 - 5/11/2019

Σε αυτό το μάθημα θα δούμε πως μπορούμε να επιταχύνουμε τον αλγόριθμο CountSketch. Υπενθύμιση: Το CountSketch είναι ένα γραμμικό σκιαγράφημα που χρησιμοποιείται για το πρόβλημα των ℓ_2 ϵ-βαρέων στοιχείων και επιτρέπει να υπολογίσεις για κάθε $i \in [n]$ έναν αριθμό x'_i ώστε

$$\|x_i - x'_i\|_2^2 \leq \epsilon \|x\|_2^2.$$

Είχαμε δεί ότι με $(\epsilon^{-1} \log n)$ χώρο και $O(n \log n)$ χρόνο μπορούμε να επιτύχουμε το ζητούμενο με πιθανότητα $1 - 1/\text{poly}(n)$. Θα δείξουμε ότι μπορούμε να κάνουμε τον χρόνο και τον χώρο $O(\epsilon^{-1} \log n \log(1/\epsilon))$. Αυτός ο χρόνος είναι εφικτός διότι δε χρειάζεται να υπολογίσουμε για κάθε $i \in [n]$ ένα x'_i, διότι για τα περισσότερα από αυτά μπορούμε να το θέσουμε ίσο με 0. Πιο συγκεκριμένα, αν $|x_i|^2 \leq \epsilon \|x\|_2^2$, το $x'_i = 0$ είναι ένας συνεπής εκτιμητής. Για την ακρίβεια, υπάρχουν το πολύ $1/\epsilon$ δείκτες $i \in [n]$ για τους οποίους πρέπει να δώσουμε ως x'_i έναν μη αριθμό μηδενικό, άρα έχει νόημα να δώσουμε μόνο τα μη μηδενικά x'_i.
Θα δείξουμε σε δύο βήματα πως να λύσουμε τα στιγμιότυπα τα οποία έχουν ακριβώς ένα μεγάλο δείκτη i^*, και στη συνέχεια θα ανάγουμε τη γενική περίπτωση σε αυτά.

Λύνοντας την 1-αραιή περίπτωση. Ας δούμε την πιο απλή περίπτωση όπου $x_{i^*} \neq 0$ και $x_i = 0, \forall i \neq i^*$. Θα χρησιμοποιήσουμε $O(\log n)$ μετρήσεις για να βρούμε το i^* και το x^*_i.
Ας πραγματοποιήσουμε τις παρακάτω μετρήσεις, ισοδύναμα ας κρατήσουμε τους παρακάτω μετρητές οι οποίοι αντιστοιχούν σε γραμμές του γραμμικού σκιαγραφήματος.

1. $y_1 = x_1 + x_3 + x_5 + \ldots$. Κρατάμε δηλαδή το άθροισμα όλων των x_i με περιττό δείκτη. Αν $y_1 = 0$, ξέρουμε ότι το i^* είναι άρτιο, αλλιώς είναι περιττό.

2. $y_2 = x_1 + x_2 + x_5 + x_6 + x_9 + x_{10} + \ldots$. Κρατάμε δηλαδή το άθροισμα όλων των x_i για τα οποία το i είναι της μορφής $4k + 1$ ή $4k + 2$. Αν $y_2 = 0$ τότε ξέρουμε ότι το i^* είναι της μορφής $4k + 2$ ή $4k + 3$. Συνδυάζοντας με το αποτέλεσμα του y_1 μπορούμε να βρούμε

αν σε ποια ακριβώς από τις δύο περιπτώσεις βρισκόμαστε. Σε κάθε περίπτωση μπορούμε να βρούμε σε ποια κλάση $mod\ 4$ βρισκμαστε.

3. Όμοια για τα πολλαπλάσια του $8, 16$ κλπ.

Πιο συμπυκνωμένα, ας θεωρήσουμε τις παρακάτω $\log n$ μετρήσεις.

$$y_b = \sum_{i\in[n]:\mathrm{bin}_b(i)=0} x_i, b \in [\log n],$$

όπου $\mathrm{bin}_b(i)$ είναι το b-οστό στοιχείο στη δυαδική αναπαράσταση του i. Αν $y_b = 0$ τότε $\mathrm{bin}_b(i^*) = 1$, αλλιώς $\mathrm{bin}_b(i^*) = 0$. Άρα μπορούμε να βρούμε κάθε δυφίο του i^*, και άρα το i^*. Προσέξτε ότι την τιμή του x_{i^*} μπορούμε να τη διαβάσουμε με άλλον έναν μετρητή που έχει το άθροισμα όλων των x_i.

Λύνοντας την περίπου 1-αραιή περίπτωση. Ας θεωρήσουμε την πιο δύσκολη περίπτωση όπου τα x_i δεν είναι απαραίτητα μηδέν, αλλά ισχύει ότι

$$|x_{i^*}|^2 \geq (1/\beta)\|x_{[n]\setminus\{i^*\}}\|_2^2,$$

για μια σταθερά β επαρκώς μεγάλη [1]. Σε αυτό το στιγμιότυπο έχουμε μια πολύ μεγάλη συντεταγμένη i^* η οποία είναι αρκετά μεγαλύτερη από όλα τα υπόλοιπα στοιχεία μαζί (σε ℓ_2 μάζα), και την οποία θέλουμε να αναγνωρίσουμε. Αυτό το πρόβλημα κουβαλάει όλη τη δυσκολία του πιο γενικού προβλήματος, και είναι ένα πρόβλημα 'αναγνώρισης μεγάλης συνιστώσας μέσα σε θόρυβο'. Η προηγούμενη λύση δε δουλεύει, και κάποιος χρειάζεται καινούρια εργαλεία για να επιτύχει το ζητούμενο. Γι αυτό το λόγο θα χρησιμοποιήσουμε τον κώδικα του Spielman στον οποίο αναφερθήκαμε στο προηγούμενο μάθημα, δηλαδή μια συνάρτηση $\mathrm{enc} : [n] \to \{0,1\}^{C\log n}$, με C κάποια απόλυτη σταθερά, η οποία ανακατασκευάζει το i αν δίνεται το $\mathrm{enc}(i)$ με λανθασμένα το πολύ $1/3$ των δυφίων του. Θα πάρουμε τις παρακάτω $2C\log n$ μετρήσεις:

$$\forall b \in [C\log n]:$$
$$y_{b,1} = \sum_{i\in[n]:\mathrm{bin}_b(\mathrm{enc}(i))=0} \sigma_b(i)x_i$$
$$y_{b,2} = \sum_{i\in[n]:\mathrm{bin}_b(\mathrm{enc}(i))=1} \sigma_b(i)x_i$$

[1] Για ένα διάνυσμα x και έν σύνολο $S \subseteq [n]$ ορίζουμε x_S το διάνυσμα που προκύπτει από το x αν μηδενίσεις κάθε $i \notin S$.

Εξήγηση: Η ιδέα είναι παρόμοια με την κατάσταση όπου δεν υπήρχε καθόλου θόρυβος, αλλά με τη βοήθεια ενός κώδικας επιδιόρθωσης σφαλμάτων. Πρώτα, 'προστατεύουμε' κάθε x_i με έναν τέτοιο κώδικα, και μετά για κάθε $b \in [C \log n]$ κρατάμε δύο μετρητές $y_{b,1}, y_{b,2}$ έτσι ώστε: στον πρώτο μετρητή να συμμετέχουν όλα τα x_i για τα οποία το $\mathrm{enc}(i)$ έχει 0 στο b-οστό του δυφίο στην δυαδική του αναπαράσταση, και στο δεύτερο μετρητή να συμμετέχουν όλα τα x_i για τα οποία το $\mathrm{enc}(i)$ έχει 1 στο b-οστό του δυφίο στην δυαδική του αναπαράσταση. Χρησιμοποιούμε και 2-ανεξάρτητα τυχαία πρόσημα $\sigma_b : [n] \to \{-1, +1\}$, τα οποία θα χρησιμεύσουν για να ελέγξουμε (ως προς την ℓ_2 νόρμα) το θόρυβο που παίρνει κάθε μετρητής.

Ο αλγόριθμος ερώτησης θα είναι ο εξής. Βρίσκω μια συμβολοσειρά $s \in \{0,1\}^{C \log n}$ ώστε το s_b να είναι 0 αν $|y_{b,1}| > |y_{b,2}|$, και 1 διαφορετικά. Αυτό μπορεί να γίνει σε $O(\log n)$ χρόνο διατρέχοντας όλα τα ζεύγη των μετρητών και κάνοντας μια σύγκριση. Χρησιμοποιώ την αντίστροφη συνάρτητη του enc πάνω στην s για να βρω έναν δείκτη i'.

Θα αποδείξουμε ότι με πιθανότητα $1 - 1/\mathrm{poly}(n)$ ο αλγόριθμος αυτός επιτυγχάνει. Αρκεί να δείξουμε ότι $i' = i^*$ με την προαναφερθείσα πιθανότητα. Ας επικεντρωθώ σε μια συγκεκριμένη τιμή του b. Ας υποθέσουμε χωρίς βλάβη της γενικότητας ότι $\mathrm{bin}_b(\mathrm{enc}(i^*)) = 0$. Τότε, με το ίδιο επιχείρημα που έχουμε δει στο AMS και στο CountSketch, έχουμε ότι

$$\mathbb{E}\left(\sum_{i \in [n] \setminus \{i^*\} : \mathrm{bin}_b(\mathrm{enc}(i)) = 0} \sigma_b(i) x_i\right) = \|x_{i \in [n] \setminus \{i^*\} : \mathrm{bin}_b(\mathrm{enc}(i)) = 0}\|_2^2 \leq \beta \|x_{i^*}\|_2^2$$

$$\mathbb{E}\left(\sum_{i \in [n] \setminus \{i^*\} : \mathrm{bin}_b(\mathrm{enc}(i)) = 1} \sigma_b(i) x_i\right) = \|x_{i \in [n] \setminus \{i^*\} : \mathrm{bin}_b(\mathrm{enc}(i)) = 1}\|_2^2 \leq \beta \|x_{i^*}\|_2^2$$

Από την ανισότητα του Markov και φράγμα ένωσης, προκύπτει ότι οι δύο ποσότητες

$$\sum_{i \in [n] \setminus \{i^*\} : \mathrm{bin}_b(\mathrm{enc}(i)) = 0} \sigma_b(i) x_i$$

$$\sum_{i \in [n] \setminus \{i^*\} : \mathrm{bin}_b(\mathrm{enc}(i)) = 0} \sigma_b(i) x_i$$

θα είναι ταυτόχρονα το πολύ $\frac{1}{16}\|x_{i^*}\|_2^2$ με πιθανότητα $1 - 32/\beta$. Στην περίπτωση που συμβαίνει αυτό το ενδεχόμενο έχουμε ότι το $y_{b,1}$ θα είναι τουλάχιστον $\|x_{i^*}\|_2 - \frac{1}{4}\|x_{i^*}\|_2 = \frac{3}{4}\|x_{i^*}\|_2$, ενώ το $y_{b,2}$ θα είναι το πολύ $\frac{1}{4}\|x_{i^*}\|_2$. Κατά συνέπεια, το s_b θα είναι ίδιο με το b-οστό δυφίο του $\mathrm{enc}(i)$ με πιθανότητα τουλάχιστον $1 - 32/\beta$.

Η πιθανότητα το s και το $\mathrm{enc}(i^*)$ να συμφωνούν σε λιγότερο από $2/3$ των δυφίων μπορεί να

χειριστεί από το φράγμα Chernoff και μπορεί να γίνει $1/\mathrm{poly}(n)$ ρυθμίζοντας το β [2]. Αυτό δείχνει ότι με τη ζητούμενη πιθανότητα θα αναγνωρίσουμε το i^*.

Ανάγοντας τη γενική περίπτωση στην περίπου 1-αραιή. Παίρνουμε $\Theta(\log(1/\epsilon)$ 2-ανεξάρτητες συναρτήσεις $h_r : [n] \to [10\beta/\epsilon]$. Κάθε τέτοια συνάρτηση 'σπάει' το $[n]$ σε $[10\beta/\epsilon]$ ομάδες. Ας κοιτάξουμε ένα συγκεκριμένο στοιχείο i^* με $|x_{i^*}|^2 \geq \epsilon\|x\|_2^2$ και μία συγκεκριμένη επανάληψη r. Η αναμενόμενη ℓ_2^2 μάζα από άλλα στοιχεία στην ομάδα του είναι

$$\mathbb{E}\left\{\sum_{i\in[n]\setminus\{i^*\}:h_r(i)=h_r(i^*)} x_i^2\right\} = \mathbb{E}\left\{\sum_{i\in[n]\setminus\{i^*\}} \delta_i x_i^2\right\} = (10/\beta)\|x_{[n]\setminus\{i^*}\|_2^2,$$

όπου εισάγαμε τυχαίες Βερνουλλι ανά 2-ανεξάρτητες μεταβλητές δ_i ώστε $\delta_i = 1$ αν $h_r(i) = h_r(i^*)$. Από Markov η ℓ^2 μάζα θα είναι το πολύ $\beta\|x_{[n]\setminus\{i^*\}}\|_2^2$ με πιθανότητα $9/10$. Άρα, αν συμβεί αυτό το ενδεχόμενο, με πιθανότητα $1 - 1/\mathrm{poly}(n)$ μπορούμε να τρέξουμε την προηγούμενη περίπτωση στο διάνυσμα $x_{\{i\in[n]:h_r(i)=h_r(i^*)\}}$ και να βρούμε το i^* με πιθανότητα $1 - 1/\mathrm{poly}(n)$. Επαναλαμβάνοντας $\Theta(\log(1/\epsilon))$ φορές, μπορούμε να εγγυηθούμε πιθανότητα αποτυχίας $\mathrm{poly}(\epsilon)$ για το i^*. Από φράγμα ένωσης η πιθανότητα να αποτύχουμε να βρούμε κάποιο ϵ-βαρύ στοιχείο είναι $(1/\epsilon) \cdot \mathrm{poly}(\epsilon)) = \mathrm{poly}(\epsilon)$. Στο τέλος θα έχουμε μια λίστα L με $O(\epsilon^{-1}\log(1/\epsilon)$ δείκτες, κάποιοι από τους οποίους ενδεχομένως επαναλαμβάνονται. Αρκεί να κρατήσουμε μία δομή CountSketch παράλληλα και να βρούμε εκτιμητές x_i' για κάθε $i \in L$.

[2] Για $b \in [C\log n]$ ορίζω τυχαία μεταβλητή $Z_b = 1$ αν s και $\mathrm{enc}(i^*)$ συμφωνούν στο b-οστό δυφίο. Τότε τα Z_b είναι ανεξάρτητες Bernoulli μεταβλητές με $\mathbb{E}Z_b \geq 1 - 32/\beta$. Η ποσότητα που αντιστοιχεί στο πλήθος των θέσεων που συμφωνούν οι δύο συμβολοσειρές είναι το $\sum_{b\in[C\log n]} Z_b$.