הרצאה 8

1 אלגוריתמים מקריים

1.1 מיון מהיר רנדומלי

אלגוריתם מיון מהיר רנדומלי - RandQS:

קלט: קבוצה $S = \{x_1 \dots x_n\}$ של מספרים ממשיים זרים.
פלט: האיברים של S בסדר ממוין.

1. אם $|S| \leq 1$, נחזיר את S.
2. נבחר ציר $p \in S$ באופן מקרי ואחיד.
3. נחלק את שאר האיברים של S לשתי קבוצות כך:
   a. $S_1 = \{x \in S : x < p\}$
   b. $S_2 = \{x \in S : x > p\}$
4. נחזיר: $\mathrm{RandQS}(S_1),\ p,\ \mathrm{RandQS}(S_2)$

משפט 1.1: זמן ריצה הצפוי הוא $\Theta(n \log n)$. בעצם, זמן הריצה הוא משתנה מקרי, וזה התוחלת שלו.

טענת עזר 1.2: נסתכל על המערך הממוין $y_1 < \cdots < y_n$, ניקח שני מספרים $y_i < y_j$.
התקיימה השוואה ביניהם אמ"מ אחד מהם היה האיבר הראשון שנבחר להיות ציר (pivot) מבין $y_i, y_{i+1} \cdots y_j$.

הוכחה: נשים לב שמתקיימת השוואה אמ"מ בזמן מסוים שניהם היו באותה קבוצה ואחד מהם נבחר בתור ציר.
יהי $y \leq k \leq j$ מספר כך ש y_k הוא המספר הראשון שנבחר בתור ציר מתוך $y_i \dots y_j$.
בהכרח קיים איבר כזה כי y_i, y_j נמצאים באותה קבוצה בהתחלה ובסוף לא.
אם k הוא i או j, תתקיים השוואה. אחרת, נשווה את שניהם ל- y_k.
מכיוון שאחד גדול ואחד קטן ממנו, הם יהיו בקבוצות נפרדות ואף פעם לא תתקיים השוואה ביניהם.

הוכחת 1.1: יהי X משתנה מקרי שסופר את המספר הכולל של השוואות. נשים לב שמספיק להראות ש $E(X) = \theta(n \log n)$.
אנחנו נוכיח ש: $E(X) = 2n \ln n + \theta(n)$.

לכל $1 \leq i < j \leq n$, יהי האינדיקטור $X_{ij} = 1$ אם השוונו בין y_i, y_j באלגוריתם. אחרת, 0.
נשים לב שלכל $1 \leq i < j \leq n$, הם מושווים לכל היותר פעם אחת (מסקנה מ1.2).
מכאן נובע שמספר ההשוואות הכולל הוא סכום האינדיקטורים של ההשוואה לכל זוג:

$$X = \sum_{1 \leq i < j \leq n} X_{ij} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}$$

(נרוץ על כל i מ-1 עד n-1 ובתוך כל i, נרוץ על כל j מ-i עד n).
ולכן מלינאריות התוחלת, נקבל ש:

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}\right) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}(X_{ij})$$

עכשיו, מטענה 1.2 נובע שמתקיים לכל $1 \leq i < j \leq n$:

$$\mathbb{E}(X_{ij}) = \mathbb{P}(X_{ij} = 1) = \frac{2}{j - i + 1}$$

(כי מתוך j-i+1 האיברים, ההשוואה תקרה רק אם בחרנו אחד מבין שני איברים). ולכן:

$$\mathbb{E}(X)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}(X_{ij})=^{\text{א}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{2}{j-i+1}=^{\text{ב}}2\sum_{i=1}^{n-1}\sum_{k=2}^{n-i+1}\frac{1}{k}=^{\text{ג}}2\sum_{k=2}^{n}\sum_{i=1}^{n-k+1}\frac{1}{k}=^{\text{ד}}2\sum_{k=2}^{n}\frac{n+1-k}{k}=$$

$$=2\sum_{k=2}^{n}\left(\frac{n+1}{k}-\frac{k}{k}\right)=2\sum_{k=2}^{n}\left(\frac{n+1}{k}\right)-2\sum_{k=2}^{n}1=$$

$$=2(n+1)\sum_{k=2}^{n}\frac{1}{k}-2(n-1)=^{\text{ה}}2(n+1)\sum_{k=1}^{n}\frac{1}{k}-2n+2-2(n+1)$$

$$=(2n+2)\sum_{k=1}^{n}\frac{1}{k}-4n$$

א – לפי מה שנובע מטענה 1.2

ב – נציב $k \coloneqq j-1+1$. ה2 יוצא החוצה.

ג – החלפת סדר סכימה: במקום לסכום לפי שורות, נסכום לפי עמודות:

	K=2	K=3	...	K=n-1	K=n
i=1	1/2	1/3	...	1 / n-1	1 / n-2
i=2	1/2	1/3	...	1 / n-1	
…	…	…	...		
i=n-2	1/2	1/3			
i=n-1	1/2				
Sum:	n-1 / 2	n-2 / 3		2 / n-1	1 / n-2

ד – השבר בתוך הסיגמא לא תלוי ב-i. תלוי רק ב-k.

ה – נרצה שהסיגמא תתחיל מ-1. אז נוריד את מה שהוספנו – $2(n+1)$.

ניזכר שבאופן כללי: $\ln n \le \sum_{k=1}^{n}\frac{1}{k} \le \ln n + 1$, כלומר הסיגמה היא $\ln n$ ועוד קבוע. אז:

$$\mathbb{E}(X)=(2n+2)\big(\ln n+\Theta(1)\big)-4n=2n\ln n+\Theta(n)+2\ln n+\Theta(1)=2n\ln n+\Theta(n)$$

כנדרש.

אינטואיציה: זה הממוצע שמצאנו. ובגלל ש $\Omega(n\ln n)$ זה חסם תחתון למיון השוואתי, זה אומר שרוב הפעמים נקבל זמן ריצה קרוב לזה. כי אם הרבה פעמים היינו הרבה מעל זה, היינו צריכים גם הרבה פעמים שהרבה מתחת לזה כדי שזה יהיה הממוצע. אבל אין זמני ריצה שהם הרבה פחות מזה (כי זה חסם תחתון) אז חייב להיות שרוב זמני הריצה קרובים לזה.

טענה 1.3: לכל קבוצה בת n מספרים ממשיים שונים, זמן הריצה יהיה $\theta(n\ln n)$ בהסתברות של לפחות $1-\frac{1}{n}$:

הוכחה: בכל ריצה של האלגוריתם, נבנה עץ בינארי שמתאר את הריצה כך:
בכל קודקוד יש תת קבוצה של S, ואיבר p. הקבוצה S היא השורש.
אם קודקוד מתאים לתת-קבוצה 'S ואיבר p , אז הילד השמאלי S_1' יהיה כל האיברים שקטנים מ-p,
והילד הימני S_2' יהיה כל האיברים שגדולים מ-p. (נניח שהקבוצות לא ריקות. אם קבוצה ריקה זה התנאי בסיס).
החלוקה של 'S לשתי קבוצות לוקחת זמן לינארי לפי גודל הקבוצה $O(|S'|)$.
כדי להוכיח את הטענה, נוכיח שגובה העץ הוא $O(\ln n)$ בהסתברות לפחות $1-\frac{1}{n}$.
נוכיח תחילה שההסתברות שהמרחק בין עלה כלשהו לשורש הוא לפחות $24\ln n$, היא לכל היותר n^{-2}:

יהי P המסלול מעלה כלשהו לשורש. קודקוד ב-P נקרא קודקוד **טוב** אם מתקיים: $\max\{|S_1'|,|S_2'|\} \le \frac{2}{3}|S'|$.
(אם הקבוצה הגדולה היא לכל היותר 2/3 מגודל הקבוצה המקורית, כלומר החלוקה היא בערך באמצע).
אחרת הוא נקרא **רע**. אנחנו נמצא חסמים עליונים לקודקודים רעים ולקודקודים טובים, וככה נגביל את אורך המסלול וגובה העץ.

טענה 1.4: עבור n מספיק גדול, לכל היותר $3\ln n$ מהקודקודים הם טובים.

הוכחה: יהיו $v_1 \cdots v_t$ הקודקודים הטובים, לפי סדר הופעתם ב-P. (1 הכי קרוב לשורש).
לכל $1 \le i \le t$, נגדיר s_i הגודל של תת-הקבוצה של S שמתאימה ל-v_i.
לכל $1 \le i \le t-1$. מתקיים (1): $s_{i+1} \le \frac{2}{3} s_i$
(הגודל של הקבוצה בכל קודקוד הוא לכל היותר 2/3 הגודל של הקודקוד הטוב הקודם במסלול). ולכן:

$$1 \le s_t \le^{א} \left(\frac{2}{3}\right)^{t-1} n \Rightarrow^{ב} \left(\frac{3}{2}\right)^{t-1} \le n \Rightarrow^{ג} t-1 \le \log_{\frac{2}{3}} n \Rightarrow t \le \log_{\frac{2}{3}} n + 1 =^{ד} \frac{\ln n}{\ln(2/3)} + 1 \le 3 \ln n$$

א – כי (1) מתקיים בכל שלב בדרך.
ב – נתעלם מ s_t. נכפול את שני האגפים ב $\left(\frac{3}{2}\right)^{t-1}$.
ג – נוציא לוג לשני האגפים.
ד – נעבור בסיס לוג.

אינטואיטיבית, כל פיצול טוב מקדם אותי המון, ולכן לא יכולים להיות הרבה כאלה (כי אם יש הרבה פשוט נסיים קודם).

עכשיו, יהי 'P ה- $24 \ln n$ קודקודים הראשונים (קרובים לשורש) במסלול. (אם P קצר יותר, פשוט ניקח את כולו).
יהי X משתנה מקרי שסופר את מספר הקודקודים הרעים ב- 'P.
לכל $u \in P'$ נגדיר אינדיקטור $X_u = 1$ אם u רע, אחרת 0.
נשים לב שמתקיים:

1) $X = \sum_{u \in P'} X_u$,
2) כל ה- X_u בת"ל,
3) $\mathbb{P}(X_u = 1) \le 2/3$.
(כי כדי שקודקוד יהיה רע, צריך לבחור מה- 2/3 איברים הקיצוניים).

בפרט, מתקיים: $\mathbb{E}(X) \le \frac{2}{3}|P'| \le 16 \ln n$.
אנחנו יודעים את התוחלת של X, והוא סכום של אינדיקטורים בת"ל אז אפשר להשתמש באי-שוויון צ'רנוף:

$$\mathbb{P}(|P| \ge 24 \ln n) =^{א} \mathbb{P}(|P'| = 24 \ln n) \le^{ב} \mathbb{P}(X \ge 21 \ln n) \le \mathbb{P}(X \ge \mathbb{E}(X) + 5 \ln n) \le^{ג} e^{-2\frac{(5 \ln n)^2}{24 \ln n}}$$
$$= e^{-2 \cdot \frac{25 \cdot (\ln n)^2}{24 \ln n}} = e^{-2 \cdot \frac{25 \ln n}{24}} = \left(\frac{1}{e^{\ln n}}\right)^{2 \cdot \frac{25}{24}} \le \frac{1}{n^2}$$

א – אם P ארוך יותר מ $24 \ln n$, אז 'P יהיה בדיוק באורך הזה.
ב – מטענה 1.4, מתוך $24 \ln n$ קודקודים, לכל היותר $3 \ln n$ יכולים להיות טובים.
ג – אי"ש צ'רנוף.

אז מכיוון שבקבוצה יש n איברים אז יש לכל היותר n עלים, חסם איחוד נותן לנו שההסתברות שיש מסלול באורך לפחות $24 \ln n$ היא לכל היותר: $n \cdot \mathbb{P}(|P| \ge 24 \ln n) \le n \cdot n^{-2} = 1/n$, כנדרש.

2 סוגי אלגוריתמים הסתברותיים

אלגוריתם לאס וגאס: הפלט תמיד נכון. זמן הריצה הוא משתנה מקרי. (לדוגמה, המיון שראינו עכשיו).
ההגדרה הסטנדרטית כוללת דרישה שהתוחלת של זמן הריצה תהיה סופית.

אלגוריתם מונטה קרלו: הפלט יכול לטעות בהסתברות (בדרך כלל קטנה מאד). יש שני תתי-סוגים:

טעות חד-צדדית: אם יצא אמת (בה"כ), זה בוודאות נכון. אבל אם יצא שקר (בה"כ) יש הסתברות לטעות.
(שני האלגוריתמים בהרצאה הקודמת הם כאלה).

טעות דו-צדדית: בכל פלט יש הסתברות לטעות.