

הרצאה 21: אלגוריתמים אקראיים

Textbook: Cormen, Leiserson, Rivest,
Stein. Introduction to Algorithms.

מרצה: אורי שטמר

בעיית התאמת המחרוזות

נתונה מחרוזת בינארית T באורך m ומחרוזת בינארית P באורך n עבור $n < m$. המטרה: למצוא את כל ההופעות של P במחרוזת T .

שימו לב: בניגוד לבעיית הקודמת עם אליס ובוב – כאן אין תקשורת. יש לנו 2 מחרוזות T, P ואנחנו רוצים למצוא את כל ההופעות של P בתוך T . המדד שיעניין אותנו כאן זה סיבוכיות החישוב.

מוטיבציה: חיפוש מילה בקובץ, או חיפוש מחרוזת ב DNA

דוגמה:

$$T = 011101001000001001$$

$$P = 0100$$

אלגוריתם נאיבי:

- לכל $1 \leq i \leq m$:
- בדוק האם המחרוזת באורך n שמתחילה בתו ה- i של T שווה ל- P

סיבוכיות: $O(n \cdot m)$

נראה אלגוריתם אקראי בסיבוכיות $O(m)$.

הערה: קיים גם אלגוריתם דטרמיניסטי בסיבוכיות כזאת שלא נראה

הרעיון: ניקח את האלגוריתם הנאיבי, אבל את הבדיקה בכל צעד (זאת שבדקת שיוויון בין מחרוזות) נחליף בפרוטוקול האקראי שראינו בפעם שעברה עבור בעיית השיוויון בין מחרוזות.

בעייה: בפעם שעברה ניתחנו את זמן הריצה וראינו שהוא $O(n)$ לכל בדיקה. לכן, אם נממש את הרעיון הנ"ל בצורה נאיבית אז נקבל אלגוריתם בסיבוכיות $O(n \cdot m)$.

פתרון: נראה איך לממש את האלגוריתם באופן יעיל יותר.

סימונים:

- $T = T[1], T[2], \dots, T[m]$ מחרוזת באורך m שנשמך כמערך
- $P = P[1], P[2], \dots, P[n]$ מחרוזת באורך n שנשמך כמערך
- נאמר ש- P מופיעה ב- T עם הזזה i אם מתקיים:
 $T[i+1] = P[1], T[i+2] = P[2], \dots, T[i+n] = P[n]$
- לכל i נסמן ב- T_i את המחרוזת באורך n שמתחילה בתו ה- $(i+1)$ של T . כלומר,
 $T_i = (T[i+1], T[i+2], \dots, T[i+n])$

האלגוריתם האקראי:

- (1) הגרל ראשוני $q < n^2 \cdot m^2$ ואתחל $S = \emptyset$
- (2) חשב $b \leftarrow P \bmod q$ (כאשר P כעל מספר בייצוג בינארי בין 0 ל- $(2^n - 1)$)
- (3) לכל $0 \leq i \leq (m - n)$ חשב:
- (א) חשב $a_i \leftarrow T_i \bmod q$ (כאשר $T_i = (T[i + 1], \dots, T[i + n])$ כמספר ביצוג בינארי)
- (ב) אם $a_i = b$ אזי $S \leftarrow S \cup \{i\}$

הערות:

- כמה ביטים יש בייצוג של q $2 \log n + 2 \log m$
- כמה ביטים יש בייצוג של T_i ושל P n
- כמה זמן לוקח חישוב ישיר של $T_i \bmod q$ $O(n)$
- כמה זמן לוקחת הבדיקה האם $a_i = b$ $O(1)$

עלינו לנתח 2 דברים לגבי האלגוריתם הזה:

1. את זמן הריצה שלו (את זה תראו בתרגול)
2. להראות שבהסתברות גבוהה האלגוריתם מחזיר בדיוק את כל מיקומי ההופעות של P ב- T

טענה 1 (שתראו בתרגול): ניתן לממש את האלגוריתם בסיבוכיות זמן $O(n + m)$ (הרעיון מאחורי טענה 1 הוא שברגע שחישבנו את a_i אז אנחנו יכולים לחשב מתוכו את a_{i+1} בצורה יעילה)

טענה 2: בהסתברות לפחות $\frac{3}{4}$, האלגוריתם מחזיר קבוצה S המכילה בדיוק את כל ההזזות של P ב- T

הוכחת טענה 2 (כלומר, ניתוח הסתברות השגיאה של האלגוריתם)

אבחנה: אם עבור i מסויים מתקיים $T_i = P$ אזי תמיד $i \in S$.

עבור i כך ש- $T_i \neq P$, נטעה ונוסיף את i ל- S אם"ם יוגרל q כך שמתקיים $T_i \bmod q = P \bmod q$ כלומר עבור i כזה נקבל שגיאה אם"ם q מחלק את $(T_i - P)$

מסקנה: S הוא פלט שגוי אם"ם קיים i כך ש- $T_i \neq P$ אבל q מחלק את $(T_i - P)$

נבחין כי $(T_i - P) < 2^n$ ונזכר בעובדה הבאה שכוחנו בשיעור שעבר:

טענה מהשיעור שעבר: עבור שלם $x \geq 55$ מתקיים

- יש ל- x לכל היותר $\log x$ מחלקים ראשוניים שונים
- ישנם לפחות $\frac{x}{2 \cdot \log x}$ מספרים ראשוניים בין 2 ל- x .

נקבע i כך ש- $T_i \neq P$. בדומה לדברים שראינו בשיעור שעבר, מתקיים:

$$\Pr[q \text{ מחלק את } (T_i - P)] \leq \frac{\log |T_i - P|}{n^2 m^2} \leq \frac{n}{n^2 m^2} = \frac{4 \log(nm)}{n \cdot m^2}$$

$$\frac{\log |T_i - P|}{2 \log(n^2 m^2)} \leq \frac{n}{2 \log(n^2 m^2)}$$

כעת, נסמן ב- i_1, i_2, \dots, i_r את כל האינדקסים עבורם $T_{i_j} \neq P$. אזי, לפי חסם האיחוד מתקיים:

$$\begin{aligned} \Pr \left[\begin{array}{c} \text{האלגוריתם} \\ \text{טועה} \end{array} \right] &= \Pr \left[\begin{array}{c} \text{קיים } j \text{ כך ש} \\ q \text{ מחלק את } (T_{i_j} - P) \end{array} \right] \\ &= \Pr \left[\left(q \text{ מחלק את } (T_{i_1} - P) \right) \text{ OR } \left(q \text{ מחלק את } (T_{i_2} - P) \right) \text{ OR } \dots \text{ OR } \left(q \text{ מחלק את } (T_{i_r} - P) \right) \right] \\ &\leq \Pr \left[q \text{ מחלק את } (T_{i_1} - P) \right] + \Pr \left[q \text{ מחלק את } (T_{i_2} - P) \right] + \dots + \Pr \left[q \text{ מחלק את } (T_{i_r} - P) \right] \\ &\leq \frac{4 \log(nm)}{n \cdot m^2} + \frac{4 \log(nm)}{n \cdot m^2} + \dots + \frac{4 \log(nm)}{n \cdot m^2} \leq m \cdot \frac{4 \log(nm)}{n \cdot m^2} = \frac{4 \log(nm)}{n \cdot m} \leq \frac{1}{4} \end{aligned}$$

(כאשר המעבר האחרון נכון עבור nm מספיק גדול)

מ.ש.ל. (טענה 2).

הקטנה הסתברות השגיאה:

אם נחזור על האלגוריתם ℓ פעמים באופן ב"ת, ההסתברות שהאלגוריתם יטעה בכל הריצות היא לכל היותר

$$\left(\frac{1}{4}\right)^\ell = 2^{-2\ell}$$

אחרת, בלפחות ריצה אחת תוחזר הקבוצה הנכונה ושאר הקבוצות יכילו את הקבוצה הנכונה.

← החיתוך הוא הקבוצה הנכונה.

זמן ריצה $O(\ell \cdot m)$.

חסמי צ'רנוף

מוטיבציה

בשתי הבעיות הקודמות שלמדנו, ראינו שיטה שטובה בהרבה מקרים להורדת הסתברות השגיאה של פרוטוקולים:

לוקחים פרוטוקול שנותן הסתברות שגיאה מסויימת ומריצים אותו הרבה פעמים

אבל בשתי הבעיות הקודמות שראינו היה לנו יתרון גדול – הטעות של הפרוטוקולים הייתה "חד כיוונית". למשל בבעיית השוויון בין מחרוזות: אם $x = y$ אז הפרוטוקול שלנו תמיד החזיר "שווים". לכן יכולנו להריץ את הפרוטוקול הרבה פעמים והספיק לנו שפעם אחת נקבל "שווים" כדי לדעת בוודאות ש- $x \neq y$.

במקרה הכללי, ישנם פרוטוקולים בהם יש הסתברות מסויימת לטעות גם אם $x = y$. פרוטוקול כזה הוא פרוטוקול בעל טעות "דו צדדית" וניתוח השגיאה עבורו נעשה (קצת) יותר מסובך. חסמי צ'רנוף יעזרו לנו (בין היתר) לנתח את ההסתברות לטעות במקרים אלו.

באופן פורמלי:

בעיית הכרעה היא בעיה בה לכל מופע x התשובה היא או "כן" או "לא". תהי A בעיית הכרעה כלשהי, עבורה לכל מופע x נסמן ב- $A(x)$ את התשובה הנכונה. לדוגמה – בעיית השיוויון בין מחרוזות. בבעיה זו המופע לבעיה מורכב מ-2 מחרוזות X, Y ומתקיים "כן" $A(X, Y) =$ אם ורק אם $X = Y$.

הגדרה: אלגוריתם אקראי M פותר בעיה A עם טעות חד כיוונית p אם מתקיים:

$$\begin{aligned} \text{לכל מופע } x \text{ כך ש- } A(x) = \text{"כן"} \text{ מתקיים ש- } \Pr[M(x) = \text{"כן"}] &= 1 \\ \text{ולכל מופע } x \text{ כך ש- } A(x) = \text{"לא"} \text{ מתקיים ש- } \Pr[M(x) = \text{"לא"}] &\geq 1 - p \end{aligned}$$

כלומר, לאלגוריתם M יש טעות חד כיוונית אם רק אחת מ-2 התשובות האפשריות עלולה להיות שגויה. לשני הפרוטוקולים האחרונים שראינו (עבור בעיית השיוויון בין מחרוזות ועבור בעיית מציאת תתי המחרוזות) הייתה טעות חד כיוונית.
(הערה: אפשר לדבר גם על אלגוריתם על טעות חד כיוונית שטועה בכיוון ההפוך...)

הגדרה: אלגוריתם אקראי M פותר בעיה A עם טעות דו כיוונית p אם מתקיים:

$$\begin{aligned} \text{לכל מופע } x \text{ כך ש- } A(x) = \text{"כן"} \text{ מתקיים ש- } \Pr[M(x) = \text{"כן"}] &\geq 1 - p \\ \text{ולכל מופע } x \text{ כך ש- } A(x) = \text{"לא"} \text{ מתקיים ש- } \Pr[M(x) = \text{"לא"}] &\geq 1 - p \end{aligned}$$

כלומר, לאלגוריתם M יש טעות דו כיוונית אם על כל קלט האלגוריתם עלול לטעות (אבל הסתברות השגיאה היא בתקווה קטנה). שימו לב, כאן אנחנו צריכים ש- $p < 1/2$ (ולא $p = 1/2$) אחרת האלגוריתם מתנהג כמו הטלת מטבע שרירותית.

תזכורת: תהי A בעיה מסויימת ויהי M אלגוריתם עם טעות חד כיוונית $1/4$. איך נוכל להקטין את הסתברות השגיאה של האלגוריתם? נריץ את האלגוריתם ℓ פעמים. אם אחת הריצות החזירה "לא" אז נחזיר "לא". אחרת נחזיר "כן". ראינו שבמקרה זה ההסתברות לטעות תהייה $4^{-\ell}$.

שאלה: תהי A בעיה מסויימת ויהי M אלגוריתם עם טעות דו כיוונית $1/4$. איך נוכל להקטין את הסתברות השגיאה של האלגוריתם? נריץ את האלגוריתם ℓ פעמים (עבור ℓ אי זוגי) ונחזיר את התשובה שתופיע יותר פעמים. שימו לב - בניגוד למקרה הקודם עכשיו לא נוכל פשוט להחזיר "לא" אם אחת הריצות החזירה "לא". מה תהייה הסתברות השגיאה עכשיו? נראה כי במקרה זה הסתברות השגיאה תהייה $c^{-\ell}$ עבור קבוע $1 < c < 4$.

מקרה פשוט: $\ell = 3$.

יהי x מופע כלשהו. ההסתברות שבריצה אחת M טועה על x היא לכל היותר $1/4$ ונניח לשם פשטות כי ההסתברות לטעות היא בדיוק $1/4$. אם נריץ את M שלוש פעמים ונחזיר את החלטת הרוב, נטעה אם"ם לפחות 2 מתוך 3 הריצות יטעו, כלומר נטעה אם 2 או 3 מתוך הריצות יטעו.

$$\text{מה ההסתברות שכל שלוש הריצות יטעו? } \left(\frac{1}{4}\right)^3 = \frac{1}{64}$$

מה ההסתברות שבדיוק 2 מתוך 3 הריצות יטעו?

$$\binom{3}{2} \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right) = \frac{9}{64}$$

כי יש $\binom{3}{2}$ דרכים לבחור את 2 המקומות שבהן M טועה, ואז יש הסתברות $\left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)$ לטעות בשני המקומות האלה ולא לטעות במקום השלישי.

סה"כ ההסתברות לטעות היא $\frac{1}{4} = \frac{8}{32} < \frac{5}{32} = \frac{10}{64}$, כלומר הקטנו את ההסתברות לטעות. אמנם לא ירדנו ל $1/64$ כמו במקרה של טעות חד כיוונית, אבל עדיין...
 (במקרה של טעות חד כיוונית, אם x הוא מופע עבורו "כן" $A(x)$, אז נשגה אם שלוש הריצות ישגו זה קורה בהסתברות $\left(\frac{1}{4}\right)^3 = \frac{1}{64}$.)

איך ננתח את ההסתברות השגיאה במקרה הכללי שבו נבצע ℓ חזרות?
 ניתן לבצע ניתוח ישיר שמרחיב את מה שעשינו עבור $\ell = 3$, אבל אנחנו נראה ניתוח כללי יותר בעזרת חסמי צ'רנוף:

משפט (חסמי צ'רנוף והופדינג):

יהיו X_1, X_2, \dots, X_n משתנים מקריים בלתי תלויים כאשר לכל i מתקיים $\Pr[X_i = 1] = p$ ו- $\Pr[X_i = 0] = 1 - p$, עבור פרמטר $0 < p < 1$. נשים לב ש-

$$\mathbb{E} \left[\sum_{i=1}^n X_i \right] = p \cdot n$$

(חסמי צ'רנוף והופדינג מראים ש- $\sum_{i=1}^n X_i$ מרוכז סביב pn)

$$(א) \text{ לכל } 0 < \delta < 1 \text{ מתקיים } \Pr[\sum_{i=1}^n X_i \geq (1 + \delta) \cdot pn] < \exp(-\delta^2 pn/4)$$

$$(ב) \text{ לכל } 0 < \delta < 1 \text{ מתקיים } \Pr[\sum_{i=1}^n X_i \leq (1 - \delta) \cdot pn] < \exp(-\delta^2 pn/4)$$

$$(ג) \text{ לכל } \delta > 0 \text{ מתקיים } \Pr[|\sum_{i=1}^n X_i - pn| \geq \delta] \leq 2 \exp(-2\delta^2/n)$$

הערות:

- קיימות גרסאות הדוקות יותר לחסמים אלו
- בדרך כלל מתייחסים לגרסה (ג) כאל "חסם הופדינג" ולגרסאות (א), (ב) כאל "חסמי צ'רנוף".

הקטנת ההסתברות השגיאה של אלגוריתם עם טעות דו כיוונית

תהי A בעייה מסויימת ויהי M אלגוריתם עם טעות דו כיוונית p עבור $0 < p < 1/2$. עבור פרמטר $\ell \in \mathbb{N}$ (אי זוגי) נגדיר אלגוריתם B אשר בהינתן קלט x מריץ ℓ פעמים (באופן "ב"ת) את $M(x)$ ומחזיר את התשובה שהופיעה הכי הרבה פעמים.

איך ניתן לחסום את ההסתברות השגיאה של אלגוריתם B כפונקציה של p, ℓ ?

נראה איך ניתן לנתח זאת בעזרת חסם הופדינג. לצורך כך נגדיר משתנים מקריים $X_1, \dots, X_\ell \in \{0,1\}$ כאשר:

$$X_i = 1 \text{ אם ורק אם ההרצה ה-} i \text{-ית של } M(x) \text{ החזירה תשובה שגויה.}$$

נשים לב ש- $\Pr[X_i = 1] \leq p$ ולכן

$$\mu = \mathbb{E} \left[\sum_{i=1}^{\ell} X_i \right] \leq p\ell$$

נזכור שאלגוריתם B עונה לפי החלטת הרוב ולכן מחזיר תשובה שגויה אם ורק לפחות $(\ell + 1)/2$ מההרצות של $A(x)$ החזירו תשובה שגויה, כלומר אם ורק אם $\sum_{i=1}^{\ell} X_i \geq (\ell + 1)/2$. לכן,

$$\begin{aligned}
\Pr[B(x) \text{ טועה}] &= \Pr\left[\sum_{i=1}^{\ell} X_i \geq \frac{\ell+1}{2}\right] = \Pr\left[\sum_{i=1}^{\ell} X_i - p\ell \geq \frac{\ell+1}{2} - p\ell\right] \\
&\leq \Pr\left[\sum_{i=1}^{\ell} X_i - p\ell \geq \ell\left(\frac{1}{2} - p\right)\right] \leq \Pr\left[\sum_{i=1}^{\ell} X_i - \mu \geq \ell\left(\frac{1}{2} - p\right)\right] \leq \Pr\left[\left|\sum_{i=1}^{\ell} X_i - \mu\right| \geq \ell\left(\frac{1}{2} - p\right)\right] \\
&\leq 2 \cdot \exp\left(-\frac{2\ell^2\left(\frac{1}{2} - p\right)^2}{\ell}\right) = 2 \cdot \exp\left(-2\ell\left(\frac{1}{2} - p\right)^2\right).
\end{aligned}$$

בפרט, עבור $p = 1/4$ נקבל שלאחר ℓ חזרות הסתברות השגיאה של B היא לכל היותר $2 \exp(-\ell/8)$.

ניתן לראות כי עבור כל הגדלה של 8 במספר החזרות, ההסתברות לשגיאה יורדת אקספוננציאלית:
עבור 8+ חזרות (לפחות 9 כי יש דרישה למספר חזרות א"ז): $2/e$
עבור 16+ חזרות: $2/e^2$
עבור 24+ חזרות: $2/e^3$
וכך הלאה...