05.01.2023 ***פרטיות מידע***

Lecture 11: The interior point problem & amplification by subsampling

Textbook: Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy

מרצה: אורי שטמר

לפני שנדבר על הנושא המרכזי שלנו היום, נראה קודם את הכלי הבא, שנותן לנו דרך להגביר את הבטחת הפרטיות של אלגוריתם נתון. נניח שיש לנו אלגוריתם $\mathcal{A}$ אשר פועל על דטהבייסים בגודל כלשהו מדומיין D ומבטיח $(1,\delta)$-פרטיות. נגדיר את האלגוריתם הבא:

אלגוריתם $\mathcal{B}$

קלט: דטהבייס $X \in D^n$

(1) בנה דטהבייס $T \subseteq X$ על ידי לקיחת כל איבר $x_i \in X$ באופן בלתי תלוי הסתברות ε

(2) החזר $\mathcal{A}(T)$

<u>משפט:</u> אם אלגוריתם $\mathcal{A}$ מקיים $(1,\delta)$-פרטיות, אז אלגוריתם $\mathcal{B}$ מקיים $(\varepsilon, \varepsilon\delta)$-פרטיות.

<u>הוכחה:</u>

נקבע מאורע F ונקבע זוג דטהבייסים שכנים $X,\quad X' = X \cup \{x'\}$.

נחשוב על הריצה של $\mathcal{B}$ על X' ונשים לב כי אם $x' \notin T$ אזי הפלט מתפלג בדיוק כמו הפלט בריצה של $\mathcal{B}$ על X.
מצד שני, אם $x' \in T$ אזי הפלט מתפלג "דומה" במובן של פרטיות דפרנציאלית, עם פרמטרים $(1,\delta)$.
כלומר,

$$\Pr[\,\mathcal{B}(X') \in F \mid x' \notin T\,] = \Pr[\,\mathcal{B}(X) \in F\,]$$

ובנוסף,

$$e^{-1} \cdot \Pr[\,\mathcal{B}(X) \in F\,] - \delta/e \;\le\; \Pr[\,\mathcal{B}(X') \in F \mid x' \in T\,] \;\le\; e \cdot \Pr[\,\mathcal{B}(X) \in F\,] + \delta$$

לכן

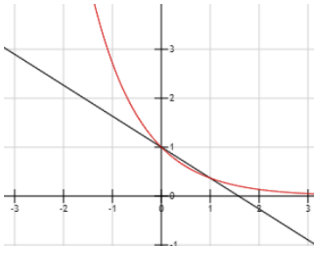
$$\begin{aligned}\Pr[\,\mathcal{B}(X') \in F\,] &= (1-\varepsilon)\cdot \Pr[\,\mathcal{B}(X') \in F \mid x' \notin T\,] + \varepsilon \cdot \Pr[\,\mathcal{B}(X') \in F \mid x' \in T\,] \\ &\le (1-\varepsilon)\cdot \Pr[\,\mathcal{B}(X) \in F\,] + \varepsilon \cdot e \cdot \Pr[\,\mathcal{B}(X) \in F\,] + \varepsilon\cdot\delta \\ &= \big(1+\varepsilon(e-1)\big)\cdot \Pr[\,\mathcal{B}(X) \in F\,] + \varepsilon\cdot\delta \\ &\le e^{2\varepsilon}\cdot \Pr[\,\mathcal{B}(X) \in F\,] + \varepsilon\cdot\delta\end{aligned}$$

הכיוון השני מתקיים באופן דומה:

$$\begin{aligned}\Pr[\,\mathcal{B}(X') \in F\,] &= (1-\varepsilon)\cdot \Pr[\,\mathcal{B}(X') \in F \mid x' \notin T\,] + \varepsilon \cdot \Pr[\,\mathcal{B}(X') \in F \mid x' \in T\,] \\ &\ge (1-\varepsilon)\cdot \Pr[\,\mathcal{B}(X) \in F\,] + \varepsilon \cdot e^{-1} \cdot \Pr[\,\mathcal{B}(X) \in F\,] - \varepsilon\cdot\delta/e \\ &= \big(1-\varepsilon(1-e^{-1})\big)\cdot \Pr[\,\mathcal{B}(X) \in F\,] - \varepsilon\cdot\delta/e \\ &\ge e^{-\varepsilon}\cdot \Pr[\,\mathcal{B}(X) \in F\,] - \varepsilon\cdot\delta/e\end{aligned}$$

(כאשר אי-השוויון האחרון נכון לכל $0 \le \varepsilon \le 1$)

מ.ש.ל.



שינויי נושא: *The interior point problem*

בעבר ראינו כי את בעיית ההיסטוגרמות אנחנו יכולים לפתור הרבה יותר טוב תחת (ε,δ)-פרטיות מאשר תחת $(\varepsilon,0)$-פרטיות (אסימפטוטית, כפונקציה של גודל הדומיין). מסתבר שהתופעה הזאת לא ייחודית להיסטוגרמות ושיש הרבה בעיות שאפשר לפתור בצורה הרבה יותר טובה כאשר $\delta>0$. היום נראה דוגמה חשובה נוספת לכך.

היזכרו בבעיית החציון:

קלט: דטהבייס $X\in D^n$ המכיל n נקודות מדומיין סדור D. אנחנו נניח כי $|D|$ הוא סופי.

המטרה: להחזיר נקודה $d\in D$ המקרבת את החציון של X במובן הזה שמספר הנקודות ב X שגדולות שוות מ d בערך זהה למספר הנקודות שקטנות שוות מ d. פורמלית, עבור נקודה $d\in D$ הגדרנו ציון

$$q_X(d)=-\left|\,|\{x\in X:x\geq d\}|-|\{x\in X:x\leq d\}|\,\right|$$

* נניח לשם פשטות שאין חזרות בדטהבייס
* נתייחס ל $-q_X(d)$ כאל "השגיאה" של d בתור קירוב לחציון

ראינו אלגוריתם לבעייה זו המשתמש במכניזם האקספוננציאלי ומחזיר פתרון עם שגיאה לכל היותר $\approx\frac{1}{\varepsilon}\log|D|$.

באופן דומה למה שראינו בהרצאות הקודמות, ניתן להראות את המשפט הבא:

משפט: לכל אלגוריתם $(\varepsilon=1\,,\delta=0)$-פרטי לבעיית החציון יש שגיאה $\Omega(\log|D|)$.

היום נראה אלגוריתם (ε,δ)-פרטי לבעייה זו המשיג שגיאה הרבה יותר נמוכה (אסימפטוטית, כפונקציה של $|D|$)

צעד ראשון: במקום לפתור את בעיית החציון, מספיק לפתור את הבעייה הפשוטה יותר הבאה:

בעיית הנקודה הפנימית *(interior point)*:

קלט: דטהבייס $X\in D^n$ המכיל n נקודות מדומיין סדור D. אנחנו נניח כי $|D|$ הוא סופי.

המטרה: להחזיר נקודה $d\in D$ המקיימת

$$\min X\leq d\leq\max X$$

נקודה d כזאת תקרא "נקודה פנימית" של הדטהבייס X.

שימו לב: בניגוד לבעיית החציון, ששם לכל פתרון אפשרי יש "איכות" מסויימת שאומרת לנו כמה הפתרון הזה קרוב לחציון, בבעיית הנקודה הפנימית אין מושג כזה של "איכות". פתרון אפשרי הוא **או נכון או שגוי**. אין באמצע ולפתרונות נכונים אין "איכות". למשל,

<u>הגדרה:</u> נאמר כי אלגוריתם $\mathcal{A}$ פותר את בעיית הנקודה הפנימית מעל דומיין D עם הסתברות שגיאה β וגודל דטהבייס n אם לכל $m \geq n$ ולכל דטהבייס $X \in D^m$ מתקיים

$$\Pr[\min X \leq \mathcal{A}(X) \leq \max X] \geq 1 - \beta$$

לשם פשטות, בהמשך ההרצאה אנו נתעלם מהסתברות הכשלון β. כלומר נניח לשם פשטות כי $\beta = 0$.

<u>שאלה:</u> למה מספיק לתכנן אלגוריתם לבעיית הנקודה הפנימית?
<u>תשובה:</u> כי יש רדוקציה פשוטה מבעיית החציון לבעיית הנקודה הפנימית. כלומר, אם תתנו לי אלג' פרטי לבעיית הנקודה הפנימית, אני יכול לבנות ממנו בקלות אלג' פרטי לבעיית החציון.

<u>רדוקציה מבעיית החציון לבעיית הנקודה הפנימית</u>

נניח: יהי $\mathcal{A}$ אלג' (ε, δ)-פרטי לבעיית הנקודה הפנימית מעל דומיין D כלשהו עם גודל דטהבייס n.
נבנה: אלג' $\mathcal{B}$ לבעיית החציון

אלגוריתם $\mathcal{B}$
<u>קלט:</u> דטהבייס $X \in D^m$
(1) מיין את X וסמן ב- $\hat{X}$ דטהבייס המכיל את n הנקודות האמצעיות ב- X
(2) הרץ $\mathcal{A}(\hat{X})$ והחזר את הפלט שלו

<u>טענה 1:</u> אלגוריתם $\mathcal{B}$ מקיים (ε, δ)-פ"ד.

<u>הוכחה:</u>
נקבע זוג דטהבייסים שכנים X, X' מאורע $F \subseteq D$.
עלינו להראות כי

$$\Pr[\mathcal{B}(X) \in F] \leq e^{\varepsilon} \cdot \Pr[\mathcal{B}(X') \in F] + \delta$$

לצורך כך, נסמן ב $\hat{X}, \hat{X}'$ את הדטהבייסים המתקבלים בשלב (1) בריצה על X, X' בהתאמה.
נשים לב ש- $\hat{X}, \hat{X}'$ הם דטהבייסים <u>שכנים</u>.
לכן,

$$\Pr[\mathcal{B}(X) \in F] = \Pr\left[\mathcal{A}(\hat{X}) \in F\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{A}(\hat{X}') \in F\right] + \delta = e^{\varepsilon} \cdot \Pr[\mathcal{B}(X') \in F] + \delta$$

מ.ש.ל.

<u>טענה 2:</u> אלגוריתם $\mathcal{B}$ פתור את בעיית החציון עם שגיאה לכל היותר n

<u>הוכחה:</u>
מכיוון ש- $\hat{X}$ מכיל את n הנקודות האמצעיות מתוך דטהבייס הקלט X, אנחנו מקבלים שאם אלגוריתם $\mathcal{A}$ מחזיר נקודה פנימית של $\hat{X}$ אז הנקודה הזאת היא כמעט חציון של X (מאוזנת עד כדי $\pm n$).

שימו לב: מסקנה מיידית מהרדוקציה הזאת היא שכל אלגוריתם $(\varepsilon, 0)$-פרטי לבעיית הנקודה הפנימית חייב לפעול על דטהבייסים בגודל $\Omega(\log|D|)$ כדי להצליח (אחרת נקבל סתירה לעובדה שתחת $(\varepsilon, 0)$-פרטיות אי אפשר לפתור את בעיית החציון עם שגיאה קטנה מ $\log|D|$)

אז מה הבנו עד עכשיו? מספיק לנו לדבר על הבעיה היותר פשוטה של מציאת נקודה פנימית. עכשיו אנחנו צריכים לראות איך פותרים את הבעיה הזאת...

אלגוריתם עם סיבוכיות $\log\log|D|$ לבעיית הנקודה הפנימית

Algorithm TreeLog

קלט: דטהבייס X המכיל n נקודות מדומיין סופי וסדור D. נניח לשם פשטות שאין חזרות בדטהבייס X.

(1) יהי T = עץ בינארי מלא עם $|D|$ עלים המתאימים לדומיין D. לכל קודקוד בעץ, נגדיר את המשקל שלו להיות מספר הנקודות מתוך X השייכות לתת העץ של הקודקוד הזה.

(2) עבור כל רמה ℓ בעץ, נסמן ב $\max_\ell$ את המשקל המקסימלי של קודקוד כלשהו ברמה ℓ. (השורש של T הוא ברמה $\ell = 0$ והעלים הם ברמה $\ell = \log|D|$)
בנוסף נגדיר $\max_{1+\log|D|} = 0$.

(3) עבור כל רמה ℓ בעץ נסמן $\text{diff}_\ell = \max_\ell - \max_{\ell+1}$

(4) נבנה דטהבייס Y המכיל כל רמה ℓ עם ריבוי diff_ℓ. למשל, אם $\text{diff}_7 = 3$ אזי Y מכיל 3 עותקים של הנקודה 7. נשים לב ש- Y הוא דטהבייס מעל דומיין בגודל $\log|D|$.

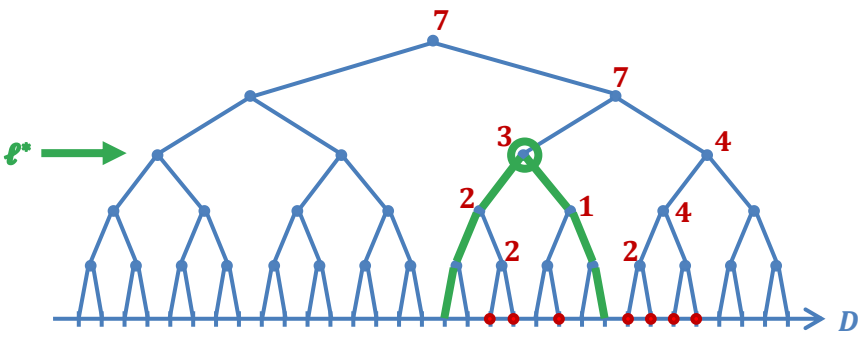
(5) נשתמש במכניזם האקספוננציאלי עם פרמטר $\varepsilon/4$ כדי למצוא קירוב לחציון של הדטהבייס Y. תהי ℓ^* הנקודה המוחזרת.

(6) נשתמש באלגוריתם (ε, δ)-פרטי להיסטוגרמות על מנת להעריך את המשקל של כל קודקוד ברמה ℓ^* של העץ T. יהי v^* קודקוד עם משקל (מוערך) גדול ביותר.

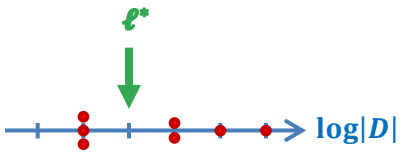
(7) נסמן ב $v^*_{\text{left-most}}$ וב- $v^*_{\text{right-most}}$ את הצאצא השמאלי ביותר והימני ביותר (בהתאמה) של v^*. בנוסף, נסמן ב v^*_{mid} עלה "אמצעי" של v^* (פורמלית, הצאצא השמאלי ביותר של הבן הימני של v^*)

(8) נעריך בעזרת המ.לפלאס את האיכות של כ"א משלושת הצאצאים האלו בתור קירוב לחציון של X ונחזיר את הטוב מבין שלושתם.

דוגמת ריצה:



max	diff
7	0
7	3
4	0
4	2
2	1
1	1



<u>משפט:</u> אלגוריתם *TreeLog* פותר את בעיית הנקודה הפנימית בהינתן דטהבייסים בגודל $n \gtrsim \log\log|D|$.

<u>הוכחה (סקיצה):</u>

נניח שבשלב (5) ההפעלה של המ.אקספ' הצליחה למצוא נקודה ℓ^* שהיא קירוב טוב לחציון של הדטהבייס Y, במובן הזה שיש לפחות $n/4$ נקודות גדולות שוות מ ℓ^* ב Y ולפחות $n/4$ נקודות קטנות שוות מ ℓ^* ב Y.

מקרה א: המשקל המקסימלי של קודקוד ברמה ℓ^* הוא לכל היותר $7n/8$.

מכיוון ש ℓ^* הוא קירוב לחציון של Y אנחנו יודעים שסך ההפרשים ברמות ℓ^* עד $\log|D|$ (כולל) הוא לפחות $n/4$. בפרט, ברמה ℓ^* חייב להיות קודקוד במשקל $n/4$, אחרת סך ההפרשים לא יכול להגיע ל $n/4$.

כלומר, ברמה ℓ^* יש קודקוד במשקל לפחות $n/4$ וכל קודקוד משקלו לכל היותר $7n/8$.

לכן הקודקוד v^* שנמצא בלשב (6) יקיים שמשקלו הוא לפחות $n/8$ ולכל היותר $7n/8$.

לכן בתת העץ המושרש בקודקוד v^* יש לפחות $n/8$ נקודות קלט ומחוץ לתת העץ הזה יש לפחות $n/8$ נקודות קלט. לכן, אחד משני הקצוות של תת העץ ($v^*_{\text{left-most}}$ או $v^*_{\text{right-most}}$) היא נקודה פנימית "עמוקה" של דטהבייס הקלט X.
(עמוקה במובן הזה שגם מימינה וגם משמאלה יש לפחות $n/16$ נקודות קלט)

מקרה ב: המשקל המקסימלי של קודקוד ברמה ℓ^* הוא לפחות $7n/8$.

המשקל של כל קודקוד אחר ברמה זו הוא לכל היותר $n/8$ ולכן בשלב (6) הקודקוד v^* שנזהה הוא בדיוק הקודקוד הכבד הזה.

מכיוון ש- ℓ^* הוא קירוב לחציון של Y, אנחנו יודעים שסך ההפרשים ברמות אפס עד ℓ^* (כולל) הוא לפחות $n/4$.

אבל מכיוון שהמשקל של v^* הוא לפחות $7n/8$ אז סך ההפרשים עד לרמה $\ell^* - 1$ הוא לכל היותר $n/8$. לכן אנו מסיקים כי $\text{diff}_{\ell^*} \geq n/8$. כלומר, ברמה הבאה, המשקל של הקודקוד v^* "מתפצל" בין שני הבנים שלו כאשר בכל בן יש משקל לפחות $n/8$. כלומר, בכל אחד משני העצים המושרשים בבנים של v^* יש לפחות $n/8$ נקודות קלט.
לכן העלה האמצעי v^*_{mid} היא נקודה פנימית "עמוקה" של דטהבייס הקלט X.
(עמוקה במובן הזה שגם מימינה וגם משמאלה יש לפחות $n/8$ נקודות קלט)

כלומר, בכל מקרה, אחת מתוך שלושת העלים שאנחנו מגדירים בשלב (7) מהווה נקודה פנימית עמוקה, ולכן נזהה אחת כזאת בעזרת המ.לפלאס (בהסתברות גבוהה).

מ.ש.ל.

<u>משפט:</u> אלגוריתם *TreeLog* מקיים פרטיות דפרנציאלית

<u>הוכחה (סקיצה):</u>

דיי ברור שצעדים *(8),(6)* משמרים פרטיות, כי בכל אחד מהם אנחנו מפעילים אלג' שאנחנו יודעים שהוא פרטי.

הנקודה העדינה היא להבין מה קורה בשלב (5). גם בשלב הזה אנחנו מפעילים אלג' פרטי, אבל לא על הדטהבייס המקורי אלא על דטהבייס אחר Y. נזכור שכאשר אנחנו מריצים את המ.אקפס כדי למצוא נקודה פנימית מתוך Y אנחנו משתמשים בפונקציית הציון

$$q_Y(\ell) = -\left| |\{y \in Y : y \geq \ell\}| - |\{y \in Y : y \leq \ell\}| \right|$$

כפי שאנחנו יודעים, הרגישות של הפונקציה הזאת ביחס לשינויי של נקודה אחת ב Y היא 2. אבל אנחנו צריכים להבין את הרגישות כפונקציה של שינויי נקודה אחת בדטהבייס המקורי X.

אז ננתח את הרגישות של $|\{y \in Y : y \leq \ell\}|$. הרגישות של הביטויי השני היא אותו דבר ואז הרגישות של $q_Y(\ell)$ תהייה פי 2 מהרגישות שנחשב.

$$|\{y \in Y : y \leq \ell\}| = \mathrm{diff}_0 + \mathrm{diff}_1 + \cdots + \mathrm{diff}_\ell$$

$$= (\max_0 - \max_1) + (\max_1 - \max_2) + \cdots + (\max_\ell - \max_{\ell+1})$$

$$= \max_0 - \max_{\ell+1}$$

כעת, שינויי של נקודה אחת בדטהבייס המקורי X יכול לשנות כ"א מהמקסימומים האלה בלכל היותר אחד, ולכן יכול לשנות את ההפרש בלכל היותר 2.
לכן הרגישות של $q_Y(\cdot)$ היא 4 (הפרש של שני ביטויים עם רגישות 2).

מכיוון שבשלב *(5)* אנחנו מריצים את המ.אקספ עם פרמטר $\varepsilon/4$ אנו מקבלים ששלב זה משמר ε-פ"ד.

סה"כ כל האלגוריתם משמר $\big(O(\varepsilon), O(\delta)\big)$-פרטיות לפי קומפוזיציה.

<u>תרגיל כיתה:</u> תכננו אלגוריתם פרטי לבעיית הנקודה הפנימית הפועל על דטהבייסים בגודל $|D|$ $n \gtrsim \log\log\log|D|$

השאלה המתבקשת כאן היא עד איפה אפשר להגיע. האם אפשר לקבל אלגוריתם (ε, δ)-פרטי עם גודל דטהבייס בלתי תלוי ב $|D|$?

<u>משפט (לא הוכחה):</u> קיים אלגוריתם (ε, δ)-פרטי לבעיית הנקודה הפנימית מעל דומיין $|D|$ עם גודל דטהבייס

$$n = \tilde{O}\left(\frac{1}{\varepsilon} \cdot \log^2\left(\frac{1}{\delta}\right) \cdot \log^*|D|\right)$$

תזכורת: הפונקציה $\log^*(\cdot)$ מחזירה את מספר הפעמים שיש להפעיל את פוקציית הלוגריתם עד שהתוצאה קטנה שווה ל 1. פורמלית:

$$\log^* y = \begin{cases} 0 & , \text{ if } y \le 1 \\ 1 + \log^*(\log y) & , \text{ if } y > 1 \end{cases}$$

זאת פונקציה שגדלה <u>מאוד מאוד</u> לאט כאשר y גדל (אבל עדיין גדלה לאינסוף כאשר $y = \infty$). ספציפית,

y	$\log^* y$
$(-\infty, 1]$	**0**
$(1, 2]$	**1**
$(2, 4]$	**2**
$(4, 16]$	**3**
$(16, 65536]$	**4**
$(65536, 2^{65536}]$	**5**
$(2^{65536}, 2^{2^{65536}}]$	**6**

<u>משפט:</u> יהיו $\varepsilon \le 1$, $\delta \le \frac{1}{4n^2}$. כל אלגוריתם (ε, δ)-פרטי לבעיית הנקודה הפנימית מעל דומיין D דורש גודל דטהבייס

$$n = \Omega(\log^* |D|)$$

איך נוכיח משפט כזה?

אנחנו נוכיח את המשפט הבא:

<u>משפט:</u> לכל n קיים דומיין D_n והתפלגות $\mathcal{P}_n$ מעל דטהבייסים בגודל $n = O(\log^* |D_n|)$ מתוך D_n כך שכל אלגוריתם פרטי נכשל במציאת נקודה פנימית מתוך $X \sim \mathcal{P}_n$ בהסתברות לפחות $1/4$.

<u>הוכחה (סקיצה, עם קצת שקרים):</u> באינדוקציה על n

<u>בסיס האינדוקציה:</u> $n = 1$
נגדיר $D_1 = \{1,2,\dots,10\}$ ונגדיר את $\mathcal{P}_1$ להיות ההתפלגות האחידה על פני D_1.
כעת, לכל אלגוריתם פרטי $\mathcal{A}$ הפועל על דטהבייסים בגודל 1 מתקיים

$$\Pr_{\substack{x\sim\mathcal{P}_1 \\ \mathcal{A}(x)}}\left[\begin{matrix}\mathcal{A}(x) \text{ מצליח במציאת} \\ \text{נקודה פנימית}\end{matrix}\right] = \Pr_{\substack{x\sim\mathcal{P}_1 \\ \mathcal{A}(x)}}[\mathcal{A}(x) = x] \le e^{\varepsilon} \cdot \Pr_{\substack{x\sim\mathcal{P}_1 \\ \mathcal{A}(1)}}[\mathcal{A}(1) = x] + \delta \le e^{\varepsilon} \cdot \frac{1}{10} + \delta < \frac{3}{4}$$

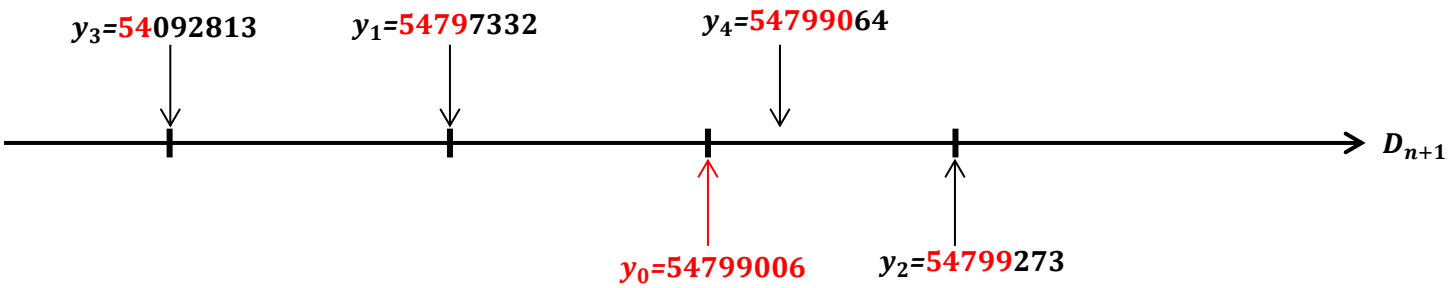
<u>צעד האינדוקציה:</u> נניח כי עבור n מסויים קיים דומיין D_n והתפלגות "קשה" $\mathcal{P}_n$ מעל דטהבייסים בגודל n מתוך D_n. נגדיר דומיין D_{n+1} והתפלגות $\mathcal{P}_{n+1}$ מעל דטהבייסים בגודל $(n+1)$ מתוך D_{n+1} באופן הבא:

הגדרת הדומיין הבא: נגדיר $D_{n+1} = \{0,1,2,\dots,10^{|D_n|} - 1\}$. אנחנו נחשוב על D_{n+1} כעל אוסף כל המחרוזות באורך $|D_n|$ בבסיס עשרוני.

הגדרת ההתפלגות הבאה $\mathcal{P}_{n+1}$:

- הגרל $(x_1, \dots, x_n) \sim \mathcal{P}_n$
- הגרל מחרוזת $y_0 \in D_{n+1}$ בהתפלגות אחידה
- עבור $i \in [n]$ נגריל מחרוזת אקראית $y_i \in D_{n+1}$ שמסכימה עם y_0 על הספרות השמאליות x_i
- נחזיר את $(y_0, y_1, \dots, y_n)$

דוגמה: $x_1=4,\ x_2=5,\ x_3=2,\ x_4=6$



מה קרה פה?
התחלנו מהתפלגות $\mathcal{P}_n$ מעל דטהבייסים בגודל n והגדרנו התפלגות $\mathcal{P}_{n+1}$ מעל דטהבייסים בגודל $n+1$

מטרה נוכחית: להראות שאם $\mathcal{P}_n$ היא התפלגות "קשה" אז גם $\mathcal{P}_{n+1}$ היא התפלגות "קשה".

איך נראה את זה? נניח בשלילה שקיים אלגוריתם פרטי $\mathcal{A}$ שפותר את $\mathcal{P}_{n+1}$, כלומר מוצא נקודה פנימית (בה"ג) כאשר מריצים אותו על דטהבייס שנדגם מתוך $\mathcal{P}_{n+1}$. נבנה אלגוריתם פרטי $\mathcal{B}$ שפותר את ההתפלגות הקשה $\mathcal{P}_n$ (וזה אנחנו מניחים שלא יכול להיות ולכן נקבל סתירה).

Algorithm $\mathcal{B}$

<u>קלט:</u> דטהבייס $X=(x_1,x_2,\dots,x_n)$ המכיל n נקודות מהדומיין D_n.

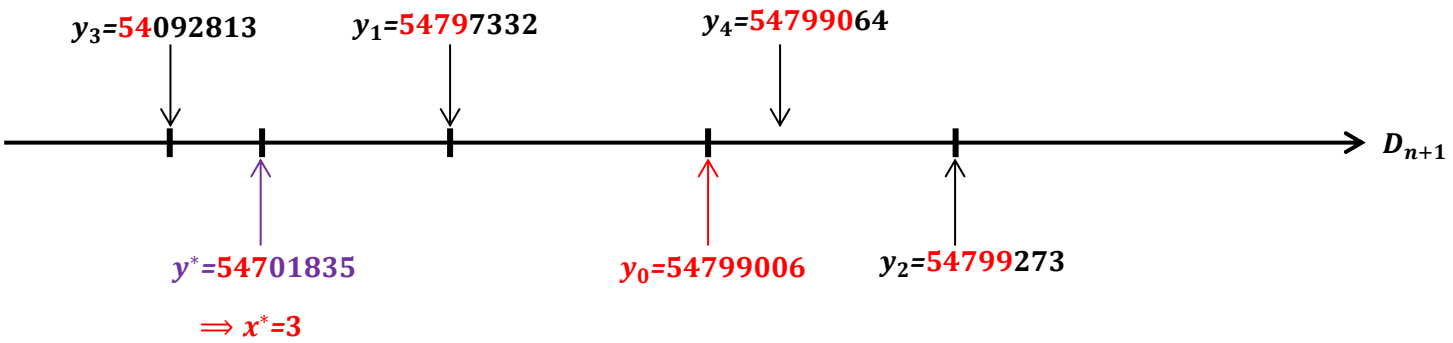
(1) הגרל $Y=(y_0,y_1,y_2,\dots,y_n)$ כמו בהגדרה של ההתפלגות $\mathcal{P}_{n+1}$ כאשר מתחילים מהדטהבייס X.

(2) הרץ $\mathcal{A}(Y)$ וקבל נקודה y^*.

(3) הגדר $x^*=$ אורך התחילית הארוכה ביותר המשותפת ל- y^* ול- y_0.

(4) החזר את x^*

בדוגמה שלנו: $x_1=4,\ x_2=5,\ x_3=2,\ x_4=6$



תחילה נשים לב שאלגוריתם $\mathcal{B}$ מקיים פ"ד. ספציפית, שינויי של נקודה אחת ב X משפיע על נקודה אחת בלבד ב Y והשינויי הזה "מוסתר" על ידי אלגוריתם $\mathcal{A}$.

מדוע הנקודה המוחזרת x^* היא נקודה פנימית של X?

- הנקודה x^* לא יכולה להיות קטנה יותר מ $\min X$ כי אז $y^* \notin [\min Y\ , \max Y]$
- הנקודה x^* לא יכולה להיות גדולה יותר מ $\max X$ מכיוון ש- $\mathcal{A}$ הוא פרטי ולכן y^* לא יכול להיות תלי יותר מדי ב y_0. באופן קצת יותר פורמלי, $(y_1, \dots, y_n)$ לא מכיל אינפורמציה לגבי הספרה מספר $(1 + \max X)$ של y_0. לכן, <u>כל אלגוריתם</u> (ובפרט אלגוריתם $\mathcal{A}$) אם מריצים אותו על קלט $(y_1, \dots, y_n)$ אז ההסתברות שהוא מנחש את הספרה הבאה ב y_0 היא לכל היותר $1/10$. לכן, מכיוון ש $\mathcal{A}$ מקיים פרטיות, אז גם כשמריצים אותו על $Y = (y_0, y_1, \dots, y_n)$ ההסתברות שהוא מנחש את הספרה הבאה ב y_0 היא קטנה.

המסקנה: אם $\mathcal{A}$ הוא אלגוריתם פרטי שפותר את ההתפלגות $\mathcal{P}_{n+1}$ אזי אלגוריתם $\mathcal{B}$ שבנינו הוא אלגוריתם פרטי שפותר את ההתפלגות $\mathcal{P}_n$.
סתירה.

<u>סיכום של מה שלמדנו לגבי בעיית הנקודה הפנימית:</u>

1. תחת $(\varepsilon, 0)$-פרטיות מספיק והכרחי גודל דטהבייס $n = \Theta(\log|D|)$
2. תחת (ε, δ)-פרטיות גודל המדגם הדרוש הוא $n = \widetilde{\Theta}(\log^*|D|)$
3. הבעיה הזאת שקולה לבעיית החציון ולכן החסמים הנ"ל מתרגמים לחסמים על השגיאה בפתרון בעיית החציון