3.11.2022 ***פרטיות מידע***

הרצאה 2: פרטיות דיפרנציאלית - תכונות

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

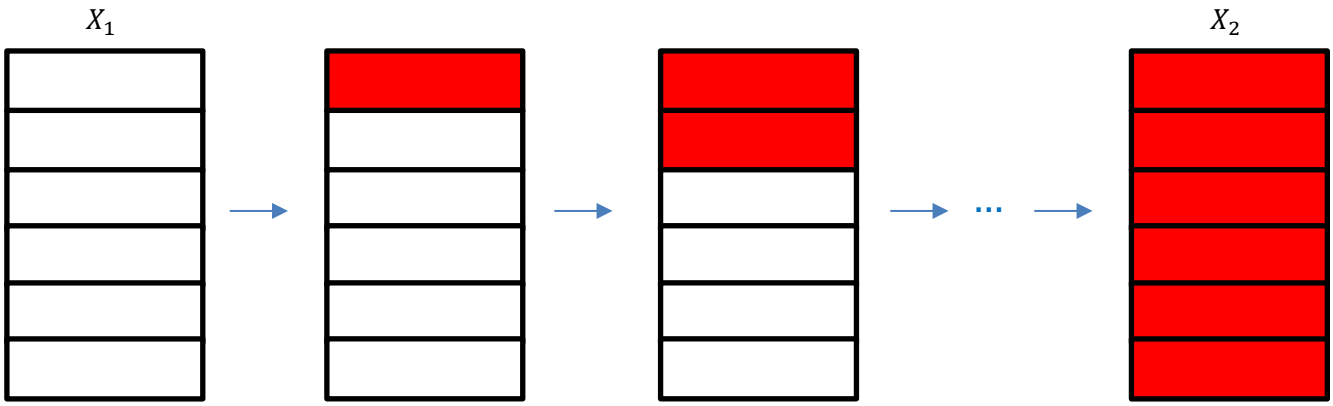
מרצה: אורי שטמר

<u>תרגיל:</u> הראו כי לכל אלגוריתם דטרמניסטי A מתקיים: או שהאלגוריתם לא משמר פרטיות דפרנציאלית (לכל ε, δ סבירים) או שהאלגוריתם לא תלוי בקלט שלו (כלומר לא תלוי בדטהבייס).

<u>פתרון:</u>
אם אלגוריתם דטרמניסטי A כן תלוי בקלט אזי ישנם שני דטהבייסים, נסמנם X_1, X_2 (לא בהכרח שכנים) כך שמתקיים

$$A(X_1) = a_1 \quad \neq \quad a_2 = A(X_2)$$

נראה שזה אומר שקיימים שני דטהבייסים <u>שכנים</u> עליהם A מתנהג אחרת. לצורך כך, "נתחיל" מ X_1 ובכל פעם נשנה שורה אחת של X_1 בשורה המתאימה מ- X_2:



אנו יודעים ש- $A(X_1) \neq A(X_2)$ ולכן חייב להיות זוג דטהבייסים <u>שכנים</u> בסדרה $\tilde{X}, \tilde{X}'$ כך ש- $A(\tilde{X}) \neq A(\tilde{X}')$.
נסמן $A(\tilde{X}) = \tilde{a}$. אזי מתקיים

$$\Pr[A(\tilde{X}) = \tilde{a}] = 1 \qquad \Pr[A(\tilde{X}') = \tilde{a}] = 0$$

ולכן עבור $\delta < 1$ אנו מקבלים ש- A לא מקיים פ"ד (לאף בחירה של ε)

<u>תכונות של ההגדרה:</u>

<u>תכונה ראשונה – *post processing*</u>
ננית ש- $M: D^n \to R$ מקיים (ε, δ)-פ"ד ותהי $A: R \to R'$. אזי גם המכניזם $A(M(\cdot))$ מקיים (ε, δ)-פ"ד.

<u>הוכחה (עבור A דטרמניסטי):</u>
יהיו X, X' שכנים (שונים בבדיוק כניסה אחת), ותהי $S \subseteq R'$.
עלינו להראות שמתקיים

$$\Pr[A(M(X)) \in S] \leq e^{\varepsilon} \cdot \Pr[A(M(X')) \in S] + \delta$$

נסמן

$$B = \{s \in R : A(s) \in S\}$$

ונקבל

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

מ.ש.ל.

מה שהטענה הזאת אומרת לנו זה שאם לקחנו דטהבייס ועשינו עליו חישוב פרטי, אז עכשיו אנחנו יכולים לקחת את תוצאת החישוב ולעשות איתה כל מה שאנחנו רוצים. לא משנה מה נעשה איתה – זה לא יפר פרטיות.

<u>תכונה שניה – *group privacy*</u>

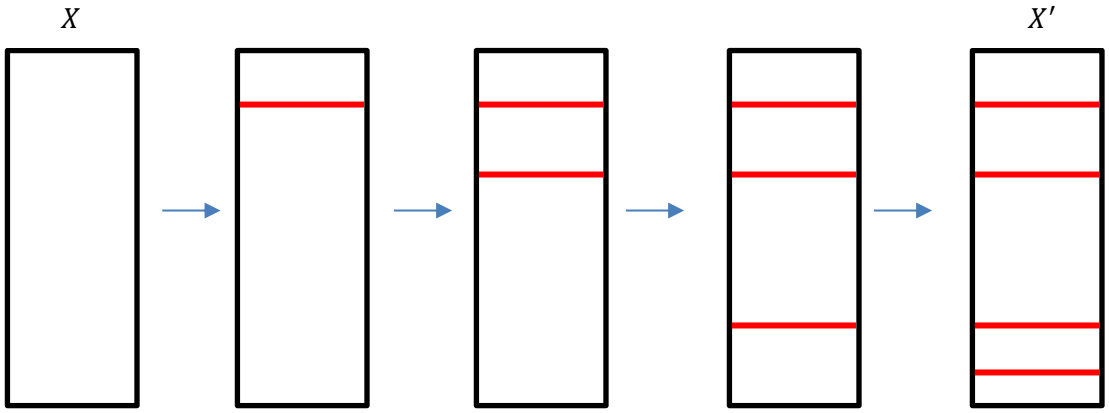
ההגדרה של פ"ד הבטיחה לי בתור בן אדם בודד שאם אתרום את המידע שלי לדטהבייס אז הפרטיות שלי תישמר כי המידע שלי כמעט לא ישפיע על תוצאת החישוב. אבל מה אם אני משפחה שלמה ששוקלת האם כדאי לנו כקבוצה לתרום את המידע שלנו?

אם $M: D^n \to R$ מקיים (ε, δ)-פ"ד אזי לכל X, X' השונים ב- t כניסות לכל היותר, ולכל $S \subseteq R$ מתקיים:

$$\Pr[M(X) \in S] \leq e^{t\cdot\varepsilon} \cdot \Pr[M(X') \in S] + t \cdot e^{t\cdot\varepsilon} \cdot \delta$$

<u>הוכחה בציור:</u>

יש לנו 2 דטהבייסים השונים ב- t כניסות. אז נפעיל את ההבטחה של פ"ד בשרשרת של t מעברים. בכל מעבר אנחנו מפסידים פקטורים שתלויים ב ε, δ. בציור, השורות האדומות הן השורות ששונות בין X, X'.



<u>הוכחה פורמלית עבור ε-פ"ד:</u>

קיימים $X_0, X_1, \dots, X_t \in D^n$ כך ש- X_i, X_{i+1} שכנים וכך ש- $X_0 = X$ וכך ש- $X_t = X'$.

עבור כל $0 \leq i < t$ מתקיים:

$$\Pr[M(X_i) \in S] \leq e^{\varepsilon} \cdot \Pr[M(X_{i+1}) \in S]$$

ולכן:

$$\Pr[M(X_0) \in S] \leq e^{\varepsilon} \cdot \Pr[M(X_1) \in S] \leq e^{\varepsilon} \cdot e^{\varepsilon} \cdot \Pr[M(X_2) \in S] \leq \cdots \leq e^{t\cdot\varepsilon} \cdot \Pr[M(X_t) \in S]$$

מ.ש.ל.

אז כל עוד $t < \frac{1}{\varepsilon}$ אנחנו מקבלים פה איזושהי הבטחת פרטיות בעלת משמעות, כי עבור $t \cdot \varepsilon$ קטן מתקיים $e^{t\varepsilon} \approx (1 + t\varepsilon)$. אבל עבור $t > \frac{1}{\varepsilon}$ יש פה גידול אקספוננציאלי...

<u>תכונה שלישית: קומפוזיציה</u>

מה קורה אם יש לי דטהבייס אחד ואני רוצה להריץ עליו 2 אלגוריתמים פרטיים?

אם M_1 מקיים $(\varepsilon_1, \delta_1)$-פ"ד ו- M_2 מקיים $(\varepsilon_2, \delta_2)$-פ"ד אזי (M_1, M_2) מקיים $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-פ"ד.

* כאן (M_1, M_2) מוגדר באופן הבא:
קלט: דטהבייס X
(1) חשב $y_1 \leftarrow M_1(X)$
(2) חשב $y_2 \leftarrow M_2(X)$
(3) החזר (y_1, y_2)

בפרט, אם M מקיים פ"ד ואני אריץ אותו על אותו דטהבייס הרבה פעמים אז לאט לאט הפרטיות תדרדר.

<u>הוכחה עבור ε-פ"ד</u> (ההוכחה עבור (ε, δ)-פ"ד משמעותית קשה יותר)

<u>נסמן:</u>

$$M_1 : D^n \to R_1$$
$$M_2 : D^n \to R_2$$

* הנחה מפשטת: R_1, R_2 בנות מניה

יהיו X, X' שכנים ויהיו $s_1 \in R_1, \ s_2 \in R_2$.
נחשב:

$$\Pr[(M_1, M_2)(X) = (s_1, s_2)] = \Pr[M_1(X) = s_1] \cdot \Pr[M_2(X) = s_2]$$

$$\leq e^{\varepsilon_1} \cdot \Pr[M_1(X') = s_1] \cdot e^{\varepsilon_2} \cdot \Pr[M_2(X') = s_2]$$

$$\leq e^{\varepsilon_1 + \varepsilon_2} \cdot \Pr[(M_1, M_2)(X') = (s_1, s_2)]$$

מ.ש.ל.

<u>שאלה:</u> בהוכחה האחרונה הנחנו כי M_1, M_2 נקבעו מראש. האם אני יכול <u>לבחור</u> את המכניזם השני על סמך התוצאה של המכניזם הראשון?

<u>משפט:</u> יהי $M_1(\cdot)$ מכניזם המשמר $(\varepsilon_1, \delta_1)$-פ"ד ויהי $M_2(\cdot, \cdot)$ מכניזם כך שלכל פרמטר s מתקיים ש- $M_2(\cdot, s)$ משמר $(\varepsilon_2, \delta_2)$-פ"ד. אזי האלגוריתם $M_3(X) = M_2\big(X, M_1(X)\big)$ משמר $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-פ"ד.

<u>הוכחה עבור ε-פ"ד:</u>

$$\Pr\big[M_2\big(X, M_1(X)\big) = s_2\big] = \sum_{s_1} \Pr[M_2(X, s_1) = s_2] \cdot \Pr[M_1(X) = s_1]$$

$$\leq \sum_{s_1} e^{\varepsilon_2} \cdot \Pr[M_2(X', s_1) = s_2] \cdot e^{\varepsilon_1} \cdot \Pr[M_1(X') = s_1]$$

$$= e^{\varepsilon_1 + \varepsilon_2} \cdot \sum_{s_1} \Pr[M_2(X', s_1) = s_2] \cdot \Pr[M_1(X') = s_1]$$

$$= e^{\varepsilon_1 + \varepsilon_2} \cdot \Pr\big[M_2\big(X', M_1(X')\big) = s_2\big]$$

מ.ש.ל.

אלגוריתמים המקיימים פרטיות דפרנציאלית

עד עכשיו הכרנו את ההגדרה של פ"ד וראינו כל מני תכונות טובות שלה. עכשיו ננסה לבנות אלגוריתם שיעמוד בהגדרה.

דוגמה: נניח שהדומיין של השורות בדטהבייס הוא בינארי: $D = \{0,1\}$. למשלף לכל אדם יש לנו ביט שאומר אם יש לו מחלה מסויימת. אז הדטהבייס שלנו הוא $X \in \{0,1\}^n$. נניח שאנחנו רוצים להעריך את $f(X) = \sum_{i=1}^{n} x_i$, כלומר רוצים להעריך את מספר האנשים שהביט שלהם הוא 1.

- בדיקת שפיות: אולי פשוט נחשב ונחזיר את $f(X)$?

ברור שזה לא עובד. למשל, אם היריב מכיר את $\sum_{i=1}^{n-1} x_i$ אז בעזרת $f(X)$ הוא לומד את x_n.

תרגיל: הצעה לאלגוריתם:
קלט: דטהבייס $X \in \{0,1\}^n$
(1) הגרל $Y \in \{-5,-4,\dots,5\}$ בהתפלגות אחידה
(2) חשב והחזר $A(X) = f(X) + Y = \sum_{i=1}^{n} x_i + Y$.

הוכיחו כי האלגוריתם הנ"ל אינו משמר פ"ד (לאף ערך סביר של ε, δ).

פתרון: נתבונן בשני הדטהבייסים השכנים הבאים:

$$X = (0,0,0,\dots,0)$$
$$X' = (1,0,0,\dots,0)$$

מתקיים:

$$\Pr[A(X) = 6] = 0$$
$$\Pr[A(X') = 6] = \frac{1}{11}$$

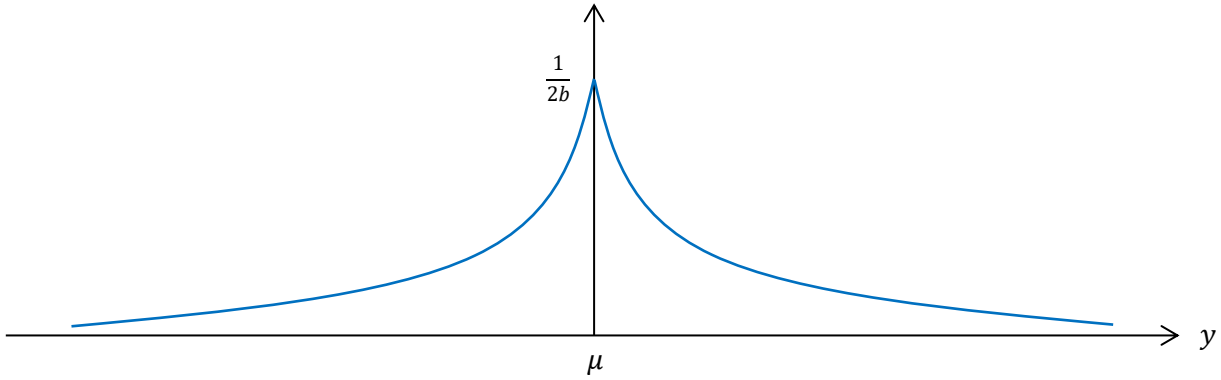
מסקנה: עבור $\delta < \frac{1}{11}$, האלגוריתם לא מקיים (ε, δ)-פ"ד לאף ערך של ε (זכרו כי אמרנו ש $\delta > 1/n$ איננו סביר).

מ.ש.ל.

הגדרה (התפלגות לפלס): למשתנה מקרי יש התפלגות $\mathrm{Lap}(\mu, b)$ אם פונקציית הצפיפות שלו היא $h_{\mu,b}(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$. עבור $\mu = 0$ נכתוב בקיצור $\mathrm{Lap}(b)$.

(תזכורת: פונקציית צפיפות של משתנה מקרי היא מתארת את צפיפות המשתנה בכל נקודה במרחב המדגם. ההסתברות שמשתנה מקרי יימצא בקטע מסוים היא האינטגרל של הצפיפות בקטע ולכן המשתנה נוטה יותר לקבל ערכים שבהם הצפיפות גבוהה.)

$$h_{\mu,b}(y) = \frac{1}{2b} \cdot \exp\left(-\frac{|y-\mu|}{b}\right)$$



<u>אבחנה:</u> אם $Y \sim \text{Lap}(\mu, b)$ ונגדיר $X = Y + k$ עבור קבוע k כלשהו, אזי $X \sim \text{Lap}(\mu + k, b)$.

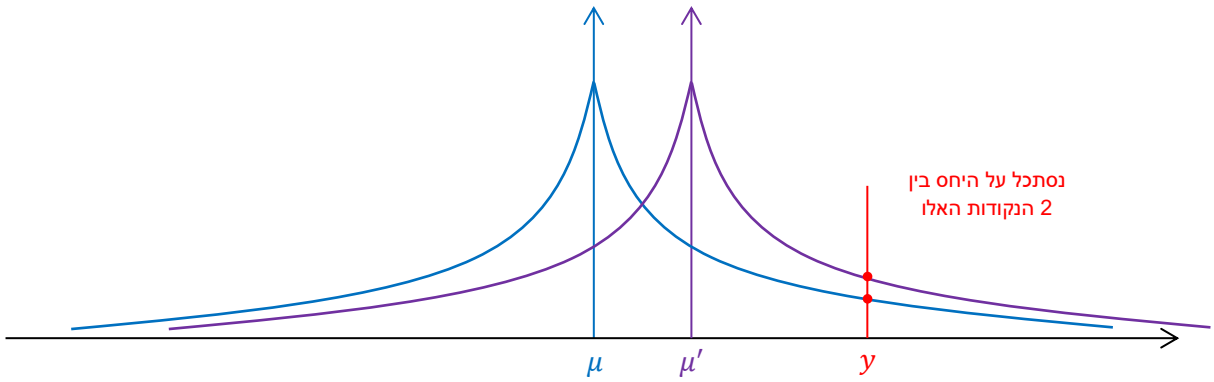
<u>הסבר:</u> זה נכון כי כאשר מוסיפים קבוע למשנה מקרי זה משנה את פונקציית הצפיפות על ידי "הזזה" של הגרף.
ספציפית, אם f_Y היא פונקציית הצפיפות של מ"מ Y ואם $X = Y + k$ אזי פונקציית הצפיפות של X היא $f_X(x) = f_Y(x - k)$. במקרה שלנו, פונקציית הצפיפות של X היא

$$f_X(x) = h_{\mu,b}(x - k) = \frac{1}{2b}\exp\left(-\frac{|x - k - \mu|}{b}\right) = h_{\mu+k,b}(x)$$

<u>טענה (תכונה של התפלגות לפלס):</u> לכל $\mu, \mu', b \in \mathbb{R}$ כך ש- $|\mu - \mu'| \leq \lambda$ ולכל $y \in \mathbb{R}$ מתקיים

$$\exp\left(-\frac{\lambda}{b}\right) \leq \frac{h_{\mu',b}(y)}{h_{\mu,b}(y)} \leq \exp\left(\frac{\lambda}{b}\right)$$

<u>הוכחה:</u>



$$\text{יחס} = \frac{h_{\mu',b}(y)}{h_{\mu,b}(y)} = \frac{\frac{1}{2b}\cdot\exp\left(-\frac{|y-\mu'|}{b}\right)}{\frac{1}{2b}\cdot\exp\left(-\frac{|y-\mu|}{b}\right)} = e^{\frac{1}{b}\underbrace{(|y-\mu|-|y-\mu'|)}_{\in[-\lambda,\lambda]}}$$

זכרו כי לפי אי-שוויון במשולש מתקיים $|y-\mu| \leq |y-\mu'| + |\mu'-\mu|$ ולכן $|y-\mu| - |y-\mu'| \leq |\mu'-\mu| \leq \lambda$ ובאותו אופן
$|y-\mu'| - |y-\mu| \leq \lambda$ כלומר $|y-\mu| - |y-\mu'| \leq |\mu'-\mu| \geq -\lambda$

סה"כ קיבלנו

$$e^{-\lambda/b} \leq \text{יחס} \leq e^{\lambda/b}$$

מ.ש.ל.

שימו לב: אם נבחר $b = 1/\varepsilon$ נקבל

$$e^{-\varepsilon\cdot\lambda} \leq \text{יחס} \leq e^{\varepsilon\cdot\lambda}$$

<u>נחזור לדוגמה שלנו:</u> הדטהבייס שלנו הוא $X \in \{0,1\}^n$ ואנחנו רוצים להעריך את $f(X) = \sum_{i=1}^{n} x_i$.

נתבונן באלגוריתם A המוגדר באופן הבא:
קלט: דטהבייס $X \in \{0,1\}^n$
(1) הגרל $Y \sim \text{Lap}\left(\frac{1}{\varepsilon}\right)$
(2) החזר $f(X) + Y$

<u>טענה:</u> אלגוריתם A מקיים ε-פ"ד.

<u>הוכחה:</u>
נקבע שני דטהבייסים שכנים $X, X' \in \{0,1\}^n$. מתקיים
$$|f(X) - f(X')| \leq 1$$
ולכן, לכל $S \subseteq \mathbb{R}$ מתקיים:

$$\Pr[A(X) \in S] = \int_S h_{f(X),\frac{1}{\varepsilon}}(y)\,dy \underbrace{\leq}_{\substack{\text{לפי התכונה} \\ \text{שראינו קודם}}} \int_S e^{\varepsilon} \cdot h_{f(X'),\frac{1}{\varepsilon}}(y)\,dy = e^{\varepsilon} \cdot \Pr[A(X') \in X]$$

מ.ש.ל.

אוקיי, אז האלגוריתם הזה מקיים $(\varepsilon, 0)$-פ"ד. כמה התשובות שהוא מחזיר מדויקות?

<u>טענה:</u> יהי $Y \sim \text{Lap}(b)$ ויהי $\Delta > 0$ כלשהו. אזי $\Pr[|Y| > \Delta] = \exp\left(-\frac{\Delta}{b}\right)$.

<u>הוכחה:</u>

$$\Pr[Y > \Delta] = \int_{\Delta}^{\infty} \frac{1}{2b} \cdot \exp\left(-\frac{y}{b}\right) dy = \frac{1}{2b} \cdot \exp\left(-\frac{y}{b}\right) \cdot (-b) \Big|_{\Delta}^{\infty} = 0 - \frac{1}{2b} \cdot \exp\left(-\frac{\Delta}{b}\right) \cdot (-b)$$

$$= \frac{1}{2} \cdot \exp\left(-\frac{\Delta}{b}\right)$$

מ.ש.ל. (הכיוון השני באופן דומה)

המסקנה היא שההסתברות שהשגיאה תהייה גדולה מ $\frac{1}{\varepsilon}$ דועכת אקספוננציאלית.

<u>סיכום עד עכשיו:</u> אם הקלט הוא $X \in \{0,1\}^n$ ואנו רוצים לחשב את $f(X) = \sum_{x \in X} x$ אזי האלג' הבא משמר ε-פ"ד:

$$\boldsymbol{f(X) + \text{Lap}\left(\frac{1}{\varepsilon}\right) = \sum_{x \in X} x + \text{Lap}\left(\frac{1}{\varepsilon}\right)} \quad \textbf{חשב והחזר}$$

<u>האינטואיציה:</u> עבור קלט שכן X' מתקיים ש- $|f(X) - f(X')| \leq 1$ והוספת רעש מהתפלגות $\text{Lap}\left(\frac{1}{\varepsilon}\right)$ מסתירה את ההבדל הזה.

<u>נוכל להכליל את זה קצת:</u> בעצם, כל מה שהניתוח שלנו הסתמך עליו זה שהערך של הפונקציה f לא משתנה הרבה בין דטהבייסים שכנים.

<u>דוגמה:</u> נתון דטהבייס X המכיל איברים מתוייגים (למשל, כל שורה בדטהבייס היא תמונה + ביט שאומר אם יש בתמונה כלב או לא). בנוסף, נתון לנו כלל תחזית h אשר בהינתן תמונה מנסה לומר אם יש בתמונה כלב או לא.

נרצה להעריך בצורה פרטית את השגיאה של h על הדטהבייס X (כלומר את מספר התמונות בדטהבייס שעליהם h טועה). נשים לב ששינוי של שורה אחת ב X יכולה לשנות בלכל היותר 1 את מספר הטעויות של h. לכן, לפי אותה אנליזה, האלגוריתם הבא משמר פרטיות:

חשב והחזר $\left[\begin{matrix}\text{מספר הטעויות}\\ \boldsymbol{X}\text{ על }\boldsymbol{h}\text{ של}\end{matrix}\right] + \mathbf{Lap}\left(\frac{\mathbf{1}}{\boldsymbol{\varepsilon}}\right)$

<u>הכללה נוספת:</u>

<u>הגדרה:</u> תהי $f: D^n \to \mathbb{R}^d$ פונקציה אשר לוקחת דטהבייס (מכיל n איברים מתוך D) ומחזירה d מספרים ממשיים. נגדיר את <u>הרגישות הגלובלית</u> של f באופן הבא:

$$\mathrm{GS}_f = \max_{\substack{X,X' \in D^n \\ \text{שכנים}}} \|f(X) - f(X')\|_1 = \sum_{j=1}^{d} \left|f_j(X) - f_j(X')\right|$$

<u>דוגמאות:</u>

- במקרה של סכום של ביטים מתקיים ש- $d = 1$ ו- $\mathrm{GS}_f = 1$
- היסטוגרמה: למשל פונקציה f לוקחת דטהבייס המכיל נתונים של אנשים, ומחזירה 10 מספרים:
  $f_1(X) =$ מספר האנשים ב X בגילאים $0 \le age < 10$
  $f_2(X) =$ מספר האנשים ב X בגילאים $10 \le age < 20$
  ...
  $f_{10}(X) =$ מספר האנשים ב X בגילאים 90 ומעלה

  אזי כאן $d = 10$ ו- $\mathrm{GS}_h = 2$

<u>משפט:</u> אם לפונקציה $f: D^n \to \mathbb{R}^d$ יש רגישות גלובלית GS_f אזי האלגוריתם הבא מקיים ε-פ"ד:

(1) הגרל $Y_1, Y_2, \dots, Y_d \sim \mathrm{Lap}\left(\frac{\mathrm{GS}_f}{\varepsilon}\right)$ באופן בלתי תלוי

(2) לכל $1 \le j \le d$ חשב והחזר $f_j(X) + Y_j$

לא נראה את ההוכחה. היא דומה להוכחה הקודמת שראינו. כמו שנראה בהמשך הקורס, הרעיון הפשוט הזה של הוספת רעש מסדר גודל של "הרגישות" של הפונקציה זהו רעיון מאוד שימושי. אבל הוא לא תמיד עובד...

<u>תרגיל לאיפוס:</u> הראו כי מתקיים:

$$\Pr\left[\mathrm{Lap}\left(\frac{4}{\varepsilon}\right) \in [a,b]\right] \le e^{\varepsilon/2} \cdot \Pr\left[\mathrm{Lap}\left(\frac{4}{\varepsilon}\right) \in [a+2, b+2]\right]$$