

## Lecture 12: Private "PAC" Learning

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

מרצה: אורי שטמר

אחד המודלים הכי בסיסיים בתאוריה של למידה נקרא *Probably Approximately Correct (PAC)* אנחנו כאן נציג גרסה מאוד פשטנית של המודל הזה.

**כללי המשחק:**

- יהי  $D$  דומיין כלשהו
- תהי  $C$  מחלקת פונקציות בינאריות מעל הדומיין  $D$ . כלומר, כל  $c \in C$  היא פונקציה  $c: D \rightarrow \{0,1\}$
- הקלט שלנו יהיה דטהבייס  $X \in (D \times \{0,1\})^n$  המכיל  $n$  נקודות מתוייגות מהדומיין  $D$ .

**סימון:** בהינתן דטהבייס  $X \in (D \times \{0,1\})^n$  ופונקציה  $h: D \rightarrow \{0,1\}$  נסמן

$$\text{error}_X(h) = \frac{1}{n} |\{(x,y) \in X : h(x) \neq y\}|$$

**המטרה:** אנחנו רוצים לתכנן אלגוריתם  $\mathcal{A}$  עם התכונות הבאות:

(1) אלגוריתם  $\mathcal{A}$  מקיים  $(\epsilon, \delta)$ -פרטיות. כלומר, לכל זוג קלטים שכנים  $X, X' \in (D \times \{0,1\})^n$  מתקיים ש-

$$\mathcal{A}(X) \approx_{\epsilon, \delta} \mathcal{A}(X')$$

(2) יהי  $X \in (D \times \{0,1\})^n$  כך שקיימת פונקצייה במחלקה  $c^* \in C$  המקיימת  $c^*(x) = y$  לכל  $(x,y) \in X$ . אזי

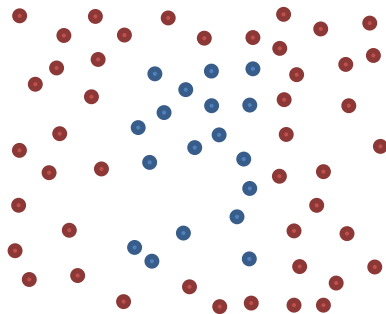
$$\Pr_{h \leftarrow \mathcal{A}(X)} [\text{error}_X(h) > \alpha] \leq \beta$$

כאשר ההסתברות כאן היא רק מעל האקראיות של  $\mathcal{A}$ .

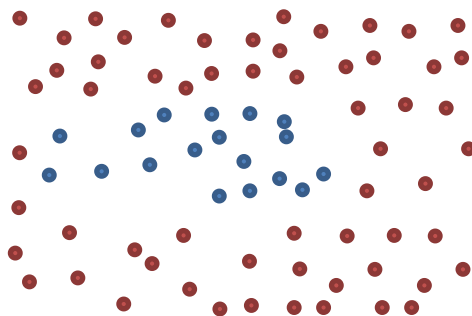
**השאלה:** מהו גודל הדטה המינימלי הדרוש על מנת שנוכל לפתור את הבעיה הנ"ל? כלומר, מהו ה  $n$  המינימלי עבורו נוכל לתכנן אלגוריתם אשר עומד בשתי הדרישות? (פרטיות תמיד ונכונות בה"ג בהנחה שהקלט תקין).

למה זאת גרסה פשטנית של המודל? כי במודל ה *PAC* "האמיתי" מניחים שיש הפלגות לא ידועה שה *data* מגיע ממנה, והמטרה היא למצוא פונקציה  $h \in C$  (נקראת השערה) עם שגיאה נמוכה ביחס להתפלגות שממנה ה *data* הגיע ולא בהכרח עם שגיאה נמוכה ביחס למדגם הספציפי שיש לנו ביד. אנחנו נתרכז בגרסה הפשטנית הנ"ל שבה המטרה שלנו היא בעצם "למידה אמפירית" כלומר אנחנו רוצים למצוא השערה עם טעות קטנה על הדטה ואנחנו לא מתעניינים בהתפלגות שה *data* הגיע ממנה אם בכלל.

**דוגמה:** נניח כי הדומיין שלנו הוא  $D = \mathbb{R}^2$  ונניח כי  $C$  היא מחלקת כל המלבנים המקבילים לצירים. במקרה כזה, הדטהבייס שלנו יכול להראות כך:



או כך:



בעצם, מה שמגדיר את "רמת הסיבוכיות" שלנו זאת ההגדרה של המחלקה  $C$ . ככל שהמחלקה הזאת יותר "מסובכת" כך החיים שלנו (אינטואיטיבית) יהיו יותר קשים.

במקרה ש-  $C$  היא מחלקת המלבנים המקבילים לצירים, הבעייה (כמו שהגדרנו אותה) היא כמעט טריוויאלית ללא אילוץ הפרטיות:

- פשוט נתעלם מכל הנקודות השליליות (האדומות בצירים הנ"ל) ונמצא את המלבן הקטן ביותר שמכיל את כל הנקודות החיוביות (הכחולות).

אבל ברגע שיש לנו את אילוץ הפרטיות, הרעיון הזה כבר לא עובד. (למה?)

### הצעה לאלגוריתם פרטי ללמידת מלבנים מקבילים לצירים במישור:

קלט: דטהבייס  $X \in (D^2 \times \{0,1\})^n$  כאשר  $D$  היא סידקרטזציה סופית כלשהי של הציר הממשי.  
כלי שנתשמש בו: אלגוריתם פרטי  $\mathcal{A}$  לבעיית הנקודה הפנימית הפועל על דטהבייסים בגודל  $m$ .

- (1) יהי  $S \subseteq X$  דטהבייס המכיל רק את הנקודות המתוייגות כ  $+1$ .
- (2) יהי  $S_1$  ההטלה של כל הנקודות ב  $S$  לציר הראשון.
- (3) נמיין את  $S_1$  נסמן ב-  $S_1^{left}$  את  $m$  הנקודות הקטנות ונסמן ב-  $S_1^{right}$  את  $m$  הנקודות הגדולות.
- (4) נחשב

$$a_1 \leftarrow \mathcal{A}(S_1^{left}) \quad b_1 \leftarrow \mathcal{A}(S_1^{right})$$

- (5) נעשה את אותו הדבר לציר השני ונקבל נקודות  $a_2, b_2$
- (6) נחזיר את המלבן המקביל לצירים המוגדר ע"י חיתוך הקטעים  $[a_1, b_1]$  הציר הראשון ו-  $[a_2, b_2]$  בציר השני.

### ניתוח פרטיות:

נניח שאלגוריתם  $\mathcal{A}$  משמר  $(\frac{\epsilon}{4}, \frac{\delta}{4})$ -פרטיות. בסה"כ אנחנו מבצעים 4 הפעלות של האלגוריתם הזה ולכן, לפי קומפוזיציה, האלגוריתם כולו משמר  $(\epsilon, \delta)$ -פרטיות.

### ניתוח השגיאה:

לשם פשטות נניח שיש "מספיק" נקודות חיוביות בדטהבייס הקלט. אחרת נוכל לזהות זאת בקלות ופשוט להחזיר את ההשערה שהיא זהותית אפס (איך?)

נשים לב שבה"ג (בהנחה שכל הריצות של  $\mathcal{A}$  הצליחו), המלבן שאנחנו מחזירים מוכל בתוך מלבן המטרה. לכן אנחנו לא טועים על נקודות שליליות, אלא רק (אולי) מפספסים חלק מהנקודות החיוביות.

כמה נקודות חיוביות אנחנו עלולים לפספס?

לפי השימוש שלנו באלגוריתם למציאת נקודה פנימית, בכל הפעלה שלו אנחנו עלולים לפספס לכל היותר  $m$  נקודות. לכן בסה"כ נטעה על לכל היותר  $4m$  נקודות שליליות.

אם נרצה שהשגיאה היחסית שלנו תהייה לכל היותר  $\alpha$  אז נדרוש:  $\frac{4m}{n} \leq \alpha$   
 או במילים אחרות, על מנת להצליח במשימה שלנו אנחנו צריכים לדרוש דטהבייס קלט בגודל לפחות  $\frac{4m}{\alpha}$ .

אנחנו יודעים מהשיעור הקודם שכדי לפתור את בעיית הנקודה הפנימית בצורה  $(\frac{\epsilon}{4}, \frac{\delta}{4})$ -פרטית מספיק לנו ש-  
 $m \approx \frac{1}{\epsilon} \cdot \log^2\left(\frac{1}{\delta}\right) \cdot \log^*|D|$  ולכן בסה"כ נקבל אלגוריתם לבעיית הקופסאות במימד 2 עם סיבוכיות מדגם של

$$n \gtrsim \frac{1}{\alpha\epsilon} \cdot \log^2\left(\frac{1}{\delta}\right) \cdot \log^*|D|$$

**תרגיל כיתה:** איך סיבוכיות המדגם הזאת משתנה אם נרצה ללמוד קופסאות מקבילות לצירים ב-  $d$  מימדים?

כלומר עכשיו:

- הקלט שלנו הוא דטהבייס  $X \in (D^d \times \{0,1\})^n$  כאשר  $D$  היא סידקרטזיציה סופית כלשהי של הציר הממשי.
- המחלקה  $C$  מכילה את כל הפונקציות ששוות אחד בתוך קופסה  $d$ -מיימדית מקבילה לצירים

הכלילו את הלומד הנ"ל למקרה זה. מהי סיבוכיות מדגם שקיבלתם?

**פורמלית:** תהי  $D \subseteq \mathbb{R}$  דיסקרטזיציה סופית כלשהי של הציר הממשי, ויהי  $d \in \mathbb{N}$ . נגדיר את מחלקת כל הקופסאות המקבילות לצירים מעל  $D^d$  באופן הבא:

$$C_{REC} = \{f_{a_1, b_1, \dots, a_d, b_d} : a_1, b_1, \dots, a_d, b_d \in D\}$$

כאשר  $f_{a_1, b_1, \dots, a_d, b_d}: D^d \rightarrow \{0,1\}$  היא הפונקציה הבאה:

$$f_{a_1, b_1, \dots, a_d, b_d}(x_1, \dots, x_d) = \begin{cases} 1 & , \forall i x_i \in [a_i, b_i] \\ 0 & , \text{else} \end{cases}$$

קלט: דטהבייס  $X \in (D^d \times \{0,1\})^n$  כאשר  $D$  היא סידקרטזיציה סופית כלשהי של הציר הממשי.  
 כלי שנתשמש בו: אלגוריתם פרטי  $\mathcal{A}$  לבעיית הנקודה הפנימית הפועל על דטהבייסים בגודל  $m$ .

1) יהי  $S \subseteq X$  דטהבייס המכיל רק את הנקודות המתוייגות כ +1.  
 2) לכל ציר  $1 \leq i \leq d$ :  
 a. יהי  $S_i$  ההטלה של כל הנקודות ב  $S$  לציר  $i$ .  
 b. נמיין את  $S_i$  ונסמן ב-  $S_i^{left}$  את  $m$  הנקודות הקטנות ונסמן ב-  $S_i^{right}$  את  $m$  הנקודות הגדולות.  
 c. נחשב

$$a_i \leftarrow \mathcal{A}(S_i^{left}) \quad b_i \leftarrow \mathcal{A}(S_i^{right})$$

3) נחזיר את המלבן המקביל לצירים המוגדר ע"י חיתוך הקטעים  $[a_i, b_i]$  בכל הצירים.

לאגל' הזה יש סיבוכיות מדגם  $\approx d^{1.5}$  (מדוע?)

**צעד אחורה:** מה מקבלים ע"י הפעלה ישירה של המכניזם האקספוננציאלי? איך נפעיל כאן את המ.אקספ?

**הקלט:** דטהבייס מתוייג  $X \in (D^d \times \{0,1\})^n$  המכיל  $n$  נקודות מתוייגות מהדומיין  $D^d$ .

**מרחב הפתרונות:** המחלקה  $C_{REC}$ . כלומר אנחנו הולכים לבחור בעזרת המ.אקספ' פונקציה אחת מתוך  $C_{REC}$ .

**פונקציית הציון:** נגדיר פונקציה  $q: (D^d \times \{0,1\})^n \times C_{REC} \rightarrow \mathbb{R}$  באופן הבא

$$q(X, f) = |\{(x, y) \in X : f(x) = y\}|$$

כלומר הציון של פונקציה  $f \in C_{REC}$  זה מספר הנקודות בדטהבייס שהיא מתייגת נכון.

**הרגישות:** הרגישות של פונקציית הציון היא 1, כי שינויי של נקודה אחת בדטהבייס משפיע בלכל היותר 1 על מספר הנקודות שכל פונקציה מתייגת נכון.

### אז מה נקבל מהפעלה של המ.אקספ' על הבעייה שלנו?

כפי שלמדנו בתחילת הקורס, המ.אקספ' משמר  $\varepsilon$ -פרטיות. בנוסף, הוא מבטיח לנו כי בהסתברות גבוהה נקבל פתרון, כלומר פונקציה  $f \in C_{REC}$ , עם ציון

$$q(X, f) \geq \text{OPT} - \frac{1}{\varepsilon} \cdot \log|C_{REC}| \approx \text{OPT} - \frac{d}{\varepsilon} \cdot \log|D|$$

בנוסף, לצורך ניתוח הנכונות אנחנו מניחים שקיימת קופסה מקבילה לצירים שמתייגת נכון את כל נקודות הקלט שלנו. במילים אחרות, אנחנו מניחים ש- $\text{OPT} = n$ . לכן, המ.אקספ' מבטיח לנו פתרון שמתייג נכון את כל הנקודות בקלט, פרט ל  $\frac{d}{\varepsilon} \cdot \log|D|$  נקודות לכל היותר. כלומר הוא מבטיח לנו פתרון עם שגיאה יחסית

$$\alpha \approx \frac{d}{\varepsilon n} \cdot \log|D|$$

לחילופין, נבודד את  $n$  מהמשוואה הנ"ל ונקבל

$$n \approx \frac{d}{\alpha \varepsilon} \cdot \log|D|$$

נשווה את הפתרון הזה לפתרון הקודם שהיה לנו:

1. הרווחנו בתלות במיימד: עכשיו סיבוכיות המדגם לניארית במיימד  $d$  כמו שאנחנו רוצים
2. הפסדנו בתלות בגודל הגריד: עכשיו סיבוכיות המדגם גדלה כמו  $\log|D|$  ולא כמו  $\log^*|D|$
3. הפסדנו את היעילות החישובית: מימוש ישיר של האלגוריתם האקספוננציאלי במקרה הזה ידרוש זמן אקספוננציאלי במיימד  $d$  בעוד שאנחנו מעוניינים באלגוריתם פולינומי

**הערה חשובה (שקצת סוטה מהנושא):** בעצם, בפתרון שעשינו כאן עם המ.אקספ' לא השתמשנו בכלל בעובדה שאנחנו מנסים ללמוד קופסאות מקבילות לצירים! כל מה שעניין אותנו כאן זה מה גודל המחלקה (ואז שילמנו לוג של זה בשגיאה). אז בעצם מה שעשינו כאן עם המ.אקספ' זאת בנייה כללית שמאפשרת ללמוד כל מחלקת פונקציות! (לא בהכרח בצורה יעילה חישובית...)

סיימנו לדבר על מלבנים מקבילים לצירים. עכשיו נדבר על מחלקה בסיסית נוספת:

**הגדרה:** יהיו  $v_1, v_2, \dots, v_d$  משתנים בולאניים (מקבלים ערך True/False). נסמן ב  $\text{CONJ}_{k,d}$  את מחלקת כל ה  $\text{conjunctions}$  (כלומר AND) של לכל היותר  $k$  ליטרלים מעל המשתנים האלו. למשל

$$f = v_2 \wedge \overline{v_5} \wedge v_6$$

פונקציה כזאת ממפה **השמה** (עבור  $d$  המשתנים) ל- 0 או 1. למשל, עבור ה  $f$  הנ"ל,  
 $f(v_1=0, v_2=1, v_3=0, v_4=1, v_5=0, v_6=1) = 1$   
 $f(v_1=0, v_2=1, v_3=0, v_4=1, v_5=1, v_6=0) = 0$

אנחנו מעוניינים לבנות לומד (פרטי) למחלקה  $\text{CONJ}_{k,d}$ . כלומר אנחנו מניחים שפונקציית המטרה היא פונקציה מתוך  $\text{CONJ}_{k,d}$ , וכל דוגמה (מתוייגת) בקלט היא השמה ל-  $d$  המשתנים (בתוספת תיוג שאומר האם תחת ההשמה הזאת פונקציית המטרה מקבלת 1 או 0).

לדוגמה, אם פונק' המטרה היא הפונק'  $f$  הנ"ל, אז הקלט שלנו יכול להראות כך:  
 $(010101,1), (010110,0), (111101,1), (111111,0), (000000,0)$

אנחנו רוצים לתכנן אלגוריתם פרטי אשר, בהנחה שהדטה מתוייג כהלכה, מחזיר השערה  $h: \{0,1\}^d \rightarrow \{0,1\}$  עם טעות קטנה על המדגם.

**באופן פורמלי:** הקלט הוא דטהבייס  $X \in (\{0,1\}^d \times \{0,1\})^n$ . יש לתכנן אלגוריתם פרטי כך שאם קיימת פונק'  $f \in \text{CONJ}_{k,d}$  עם טעות  $\text{error}_X(f) = 0$  אז בה"ג הפלט הוא השערה  $h: \{0,1\}^d \rightarrow \{0,1\}$  עם טעות

$$\text{error}_X(h) = \frac{1}{n} |\{(x,y) \in X : h(x) \neq y\}|$$

קטנה.

**הערות:**

1. האלג' צריך לשמר פרטיות בכל מקרה. כלומר דרישת הפרטיות חייבת להתקיים גם אם לא קיימת פונק'  $f \in \text{CONJ}_{k,d}$  עם טעות  $\text{error}_X(f) = 0$
2. הפונקציה שנחזיר  $h$  לא בהכרח חייבת להיות מתוך המחלקה  $\text{CONJ}_{k,d}$ . זה נחמד אם כן, אבל זה לא הכרחי. כשדיברנו על הלומד לקופסאות, הפלט שלנו היה בעצמו קופסה. אבל באופן כללי אנחנו מוכנים לקבל בתור פלט כל פונקציה, העיקר שהיא תדע לתייג נכון נקודות...

### שלב 0: מה מקבלים מהמכניזם האקספוננציאלי?

המ.אקספ' מבטיח השערה עם שגיאה לכל היותר  $\alpha$  (בהסברות גבוהה) בהינתן שגודל הדטהבייס הוא

$$n = |X| \geq \frac{1}{\epsilon\alpha} \log |\text{CONJ}_{k,d}| \approx \frac{1}{\epsilon\alpha} \log(d^k) = \frac{k}{\epsilon\alpha} \log(d)$$

זאת תוצאה טובה מאוד מבחינת גודל הדטה הדרוש. אבל הבעיה היא שמימוש ישיר של המ.אקספ' במקרה הזה הוא מאוד לא יעיל. דורש זמן  $\approx d^k$ .

איך נוכל לתכנן אלגוריתם יעיל יותר (מבחינה חישובית)?

### שלב 1: איך אפשר לפתור את הבעיה ביעילות ללא אילוצי פרטיות?

### Non-private algorithm for conjunctions

קלט: דטהבייס  $X \in (\{0,1\}^d \times \{0,1\})^n$

$$(1) \text{ אתחל } h = v_1 \wedge \bar{v}_1 \wedge v_2 \wedge \bar{v}_2 \wedge \dots \wedge v_d \wedge \bar{v}_d$$

(שימו לב כי לאחר האתחול מתקיים  $h \equiv 0$ )

(2) מחק מ  $h$  כל ליטרל "שסותר" נקודת קלט המתוייגת כ-1.  
למשל, אם  $(0110,1) \in X$  אזי נמחק את  $v_1, \bar{v}_2, \bar{v}_3, v_4$  מ- $h$ .

### דוגמת ריצה:

קלט:

$(010101,1), (010110,0), (111101,1), (111111,0), (000000,0), (101101,1), (111101,1), (000101,1)$

נאתחל את ההשערה

$$h = v_1 \wedge \bar{v}_1 \wedge v_2 \wedge \bar{v}_2 \wedge v_3 \wedge \bar{v}_3 \wedge v_4 \wedge \bar{v}_4 \wedge v_5 \wedge \bar{v}_5 \wedge v_6 \wedge \bar{v}_6$$

תחילה נשים לב שהאלג' הנ"ל מתעלם מנקודות הקלט השליליות (אלה המתוייגות כ-0).  
אז נתסכל רק על נקודות הקלט החיוביות:

$(010101,1), (111101,1), (101101,1), (111101,1), (000101,1)$

כעת, נעבור על הנקודות החיוביות האלה אחת ונתקן את  $h$  בהתאם:

- בשביל לתייג נכון את  $(010101,1)$ , ההשערה שלנו לא יכולה להכיל את הליטרלים  $v_1, \bar{v}_2, v_3, \bar{v}_4, v_5, \bar{v}_6$ .
- בשביל לתייג נכון את  $(111101,1)$ , ההשערה שלנו לא יכולה להכיל את הליטרלים  $\bar{v}_1, \bar{v}_2, \bar{v}_3, \bar{v}_4, v_5, \bar{v}_6$ .
- בשביל לתייג נכון את  $(101101,1)$ , ההשערה שלנו לא יכולה להכיל את הליטרלים  $\bar{v}_1, v_2, \bar{v}_3, \bar{v}_4, v_5, \bar{v}_6$ .
- בשביל לתייג נכון את  $(111101,1)$ , ההשערה שלנו לא יכולה להכיל את הליטרלים  $\bar{v}_1, \bar{v}_2, \bar{v}_3, \bar{v}_4, v_5, \bar{v}_6$ .
- בשביל לתייג נכון את  $(000101,1)$ , ההשערה שלנו לא יכולה להכיל את הליטרלים  $v_1, v_2, v_3, \bar{v}_4, v_5, \bar{v}_6$ .

נמחק את כל הליטרלים האלה מ- $h$  ונשאר עם:

$$h = v_4 \wedge \bar{v}_5 \wedge v_6$$

### ניתוח האלגוריתם הלא פרטי:

לצורך ניתוח הנכונות אנחנו מניחים שקיימת פונקציית מטרה  $f \in \text{CONJ}_{k,d}$  שמסבירה נכון את הדטה. תהי  $h$  ההשערה המתקבלת בסיום הריצה.

תחילה נשים לב כי  $h$  לא טועה על נקודות חיוביות, מכיוון שדאגנו למחוק מ- $h$  כל ליטרל שגורם לנקודות האלה להיות מתוייגות כ-0. נשאר להראות ש  $h$  לא טועה על נקודות שליליות.  
טענת עזר: אם מחקנו ליטרל  $\ell$  מ- $h$  אזי הליטרל הזה לא נמצא בפונקציית המטרה  $f$ .

הוכחה: מחקנו את הליטרל הזה מכיוון שקיימת נקודה חיובית שסותרת את הליטרל הזה. כל פונקציה שמכילה את הליטרל הזה מתייגת את הנקודה הזאת כ-0. אבל הנקודה מתייגת כ-1 ע"י פונקציית המטרה, ולכן הליטרל הזה לא נמצא בפונקציית המטרה.

מסקנה 1: פונקציית המטרה  $f$  מוכלת בתוך  $h$  (או ליתר דיוק, כל הליטרלים בפונקציית המטרה נמצאים ב- $h$ , ויתכן שב- $h$  יש ליטרלים נוספים).

מסקנה 2: ההשערה  $h$  שאנחנו מחזירים לא טועה על אף נקודת קלט שלילית (המתוייגת כ-0).

**הוכחה:** תהי  $x \in \{0,1\}^d$  נקודת קלט המתוייגת כ-0. כלומר פונקציית המטרה  $f$  מתייגת נקודה זו כ-0. לכן קיים לפחות ליטרל אחד ב-  $f$  שלא מסתפק תחת ההשמה  $x$ . מכיוון שהליטרל הזה קיים גם ב-  $h$ , אזי גם  $h$  מתייגת את הנקודה הזאת כ-0.

לסיכום, ההשערה שאנחנו מחזירים לא טועה על אף נקודת קלט חיובית ולא טועה על אף נקודת קלט שלילית. כלומר,

$$\text{error}_x(h) = 0$$

## שלב 2: איך נוכל לתכנן גרסה פרטית של האלגוריתם הזה?

דרך שקולה להציג את האלגוריתם (הלא פרטי) הנ"ל היא:

- נעבור על הליטרלים האפשריים אחד אחד  $\ell = v_1, \bar{v}_1, \dots, v_d, \bar{v}_d$ .
- לכל ליטרל כזה  $\ell$  נספור כמה נקודות קלט חיוביות סותרות אותו. אם המספר גדול מאפס אז נמחוק את  $\ell$  מההשערה שאנחנו בונים.

נוכל לתכנן גרסה פרטית ע"י כך שנרעיש את הספירות האלה. נקודה עדינה שאנחנו צריכים לשים לב אליה: בניתוח הנכונות הסתמכנו על העובדה שפונק' המטרה מוכלת בתוך ההשערה שאנחנו מחזירים (כדי להראות שאנחנו לא טועים על נקודות שליליות). בשביל זה הסתמכנו על העובדה שאם ליטרל מסויים לא נכנס להשערה שאנחנו בונים אז הוא גם לא מופיע בפונק' המטרה, כי הייתה לו סתירה בנתונים. כדי לשמר את התכונה הזאת אחרי שנוסיף רעש לספירות, נרצה לקחת "טווח ביטחון" כדי להחליט לוותר על ליטרל. כלומר, במקום לוותר על ליטרל אם הייתה לו לפחות סתירה אחת בנתונים, נוותר עליו אם אחרי ההרעשה היו לו "מספיק" סתירות כך שנהייה בטוחים שקיימת לו לפחות סתירה "אמיתית" אחת.

### Private algorithm for conjunctions

**קלט:** דטהבייס  $X \in (\{0,1\}^d \times \{0,1\})^n$

(1) סמן ב-  $X^1$  את הנקודות החיוביות ב-  $X$  ואתחל  $h = v_1 \wedge \bar{v}_1 \wedge v_2 \wedge \bar{v}_2 \wedge \dots \wedge v_d \wedge \bar{v}_d$

(2) עבור  $\ell = v_1, \bar{v}_1, \dots, v_d, \bar{v}_d$

(א) סמן ב-  $\#_{\ell \rightarrow 0}(X^1)$  את מספר נקודות הקלט החיוביות שהליטרל  $\ell$  לא מסתפק עליהן. כלומר זהו מספר נקודות הקלט החיוביות שסותרות את הליטרל  $\ell$

(ב) חשב  $\widehat{\#_{\ell \rightarrow 0}} = \#_{\ell \rightarrow 0}(X^1) + \text{Lap}\left(\frac{2d}{\epsilon}\right)$

(ג) אם  $\widehat{\#_{\ell \rightarrow 0}} > \frac{2d}{\epsilon} \log \frac{2d}{\beta}$  אזי מחק את הליטרל  $\ell$  מההשערה  $h$

(3) החזר את ההשערה  $h$

### ניתוח פרטיות:

לאורך הריצה ישנן  $2d$  הפעלות של המכניזם הלפליסי. לכן, לפי קומפוזיציה, אם נבצע כל הפעלה עם פרמטר פרטיות של  $\epsilon/(2d)$  אז האלגוריתם כולו ישמר  $\epsilon$ -פרטיות.

### ניתוח הדיוק:

לפי התכונות של התפלגות לפלס (+חם האיחוד), בהסתברות לפחות  $1 - \beta$  מתקיים שכל הרעשים לאורך כל הריצה הם לכל היותר  $\frac{2d}{\epsilon} \log \frac{2d}{\beta}$  בערך מוחלט. נמשיך בהוכחה בהנחה שזה המצב.

לכן, אם מחקנו ליטרל  $\ell$ , כלומר אם  $\widehat{\#_{\ell \rightarrow 0}} > \frac{2d}{\varepsilon} \log \frac{2d}{\beta}$ , אז בהכרח מתקיים  $\#_{\ell \rightarrow 0}(X^1) > 0$  כלומר ישנה לפחות דוגמה חיובית אחת בנתונים שסותרת את הליטרל  $\ell$  ולכן  $\ell$  לא מופיע בפונק' המטרה. כלומר, כמו בניתוח האלגוריתם הלא-פרטי, פונקציית המטרה מוכלת בתוך ההשערה שאנחנו מחזירים ולכן אנחנו לא טועים על דוגמאות שליליות.

על כמה דוגמאות חיוביות נוכל לטעות?

אנחנו טועים על דוגמה חיובית אם יש ליטרל שהיינו צריכים למחוק, אבל לא מחקנו. כלומר ליטרל כך ש  $\#_{\ell \rightarrow 0}(X^1) > 0$ , אבל  $\widehat{\#_{\ell \rightarrow 0}} \leq \frac{2d}{\varepsilon} \log \frac{2d}{\beta}$ . מכיוון שהנחנו שהרעשים חסומים בערך מוחלט, אז אנחנו מסיקים שמתקיים  $\#_{\ell \rightarrow 0}(X^1) \leq 2 \cdot \frac{2d}{\varepsilon} \log \frac{2d}{\beta}$ . כלומר כל ליטרל שהיינו צריכים למחוק אבל לא מחקנו, טועה על לכל היותר  $\frac{4d}{\varepsilon} \log \frac{2d}{\beta}$ . מכיוון שיש לכל היותר  $2d$  ליטרלים, אז בסה"כ נטעה על לכל היותר  $\frac{8d^2}{\varepsilon} \log \frac{2d}{\beta}$  נקודות חיוביות.

לכן, אם נרצה שמספר הטעויות הזה יהווה לכל היותר חלק  $\alpha$  ממספר נקודות הקלט, אז נדרוש ש-

$$\underbrace{\alpha n}_{\text{מספר נקודות שמוותר לטעות עליהן}} \geq \underbrace{\frac{8d^2}{\varepsilon} \log \frac{2d}{\beta}}_{\text{חסם על מספר הנקודות שטועים עליהן}}$$

כלומר נדרוש

$$n \geq \frac{8d^2}{\alpha \varepsilon} \log \frac{2d}{\beta}$$

אפשר לשפר את זה קצת עם קומפוזיציה חזקה, אבל זה בכל מקרה מאוד בזבזני ביחס לגודל הדטה הדרוש למ.אקספ'. בפרט, עכשיו  $n$  צריך להיות פולינומי ב  $d$  במקום ב  $k$ .

### שלב 3: אלגוריתם פרטי יותר יעיל

הבזבזנות בפתרון הקודם נבעה מכך שבדקנו את  $2d$  הליטרלים האפשריים אחד אחד, ("ושילמנו" על כל בדיקה כזאת). לעומת זאת, הפתרון הראשוני שראינו עם המ.אקספ בחר בבת אחת  $k$  ליטרלים "טובים" מתוך  $2d$  הליטרלים האפשריים. זה היה מצויין מבחינת גודל הדטה הדרוש, אבל יקר מבחינה חישובית, כי יש  $(2d)^k$  אפשרויות ל  $k$ -איות כאלה...

מחשבה חדשה: אולי במקום לבחור  $k$  ליטרלים טובים בבת אחת, נבחר אותם אחד אחד בצורה איטרטיבית? כלומר ננסה לתכנן אלגוריתם איטרטיבי שבכל שלב בוחר ליטרל אחד טוב ומוסיף אותו להשערה שהוא בונה.

מתי ליטרל הוא טוב?

### הרעיון:

נזכור שפונקציית המטרה מורכבת מ-  $k$  ליטרלים. נסמן את הליטרלים בפונקציית המטרה כ-  $\ell_1, \dots, \ell_k$ . מכיוון שאלו הם הליטרלים בפונקציית המטרה, ומכיוון שפ' המטרה מתייגת נכון את כל הדטה, מתקיים:

- כל נקודה חיובית  $x$  מספקת את כל הליטרלים  $\ell_1, \dots, \ell_k$ .
- לכל נקודת קלט שלילית  $x$  קיים ליטרל  $\ell_j$  שלא מסתפק תחת  $x$ .



בפרט, קיים ליטרל כלשהו  $\ell$  שמסתפק תחת כל הנקודות החיוביות וְלא מסתפק תחת לפחות חלק  $\frac{1}{k}$  מהנקודות השליליות. נתכנן אלגוריתם שבכל שלב בוחר בצורה פרטית ליטרל כזה (עם המכניזם האקספוננציאלי) ומוסיף אותו להשערה שהוא בונה.

### Private algorithm for conjunctions, take 2

קלט: דטהבייס  $X \in (\{0,1\}^d \times \{0,1\})^n$

(4) עבור  $j = 1$  עד  $J$

(א) סמן ב-  $X^0, X^1$  את הנקודות השליליות והחיוביות ב-  $X$ , בהתאמה.

(ב) עבור כל ליטרל אפשרי  $\ell \in \{v_1, \bar{v}_1, \dots, v_d, \bar{v}_d\}$ , סמן ב  $\#_{\ell \rightarrow 0}(X^1)$  את מספר נקודות הקלט החיוביות שהליטרל  $\ell$  לא מסתפק עליהן. באופן דומה, סמן ב  $\#_{\ell \rightarrow 0}(X^0)$  את מספר נקודות השליליות שהליטרל  $\ell$  לא מסתפק עליהן.

(ג) בחר עם המ.אקספ' ליטרל  $\ell \in \{v_1, \bar{v}_1, \dots, v_d, \bar{v}_d\}$  עם פונקציית ציון:

$$q(\ell) = \min \left\{ \#_{\ell \rightarrow 0}(X^0) - \frac{|X^0|}{k}, -\#_{\ell \rightarrow 0}(X^1) \right\}$$

נסמן ב-  $\ell_j$  את הליטרל המוחזר.

(ד) מחק מ-  $X$  כל נקודה  $(x, y)$  כך ש-  $\ell_j$  לא מסתפק תחת  $x$ .

$$(5) \text{ החזר את ההשערה } h = \ell_1 \wedge \ell_2 \wedge \dots \wedge \ell_j$$

### ניתוח פרטיות (סקיצה):

לאורך הריצה ישנן  $J$  הפעלות של המכניזם האקספוננציאלי. לכן, לפי קומפוזיציה, אם נבצע כל הפעלה עם פרמטר פרטיות של  $\varepsilon/J$  אז כל האלגוריתם כולו ישמר  $\varepsilon$ -פרטיות.

אז נניח שכל הפעלה של המ.אקספ מתבצעת עם פרמטר פרטיות  $\varepsilon/J$ .

### ניתוח השגיאה (סקיצה):

תחילה נשים לב כי בכל איטרציה של האלגוריתם קיים ליטרל  $\ell^*$  עם ציון  $q(\ell^*) = 0$ . הסיבה לכך היא, כמו שאמרנו קודם, שפונקציית המטרה מורכבת מ-  $k$  ליטרלים שמתייגים נכון את כל הדטה. כלומר, כל אחד מהליטרלים האלה מסתפק תחת כל נקודה חיובית בקלט, ומכיוון שכולם ביחד מתייגים נכון את כל הנקודות השליליות, אז לפחות אחד מבין הליטרלים בפונקציית המטרה מתייג נכון חלק  $1/k$  מהנקודות השליליות. לליטרל כזה יש ציון 0. (זה נשאר נכון גם אם מחקנו חלק מהדטה...)

לכן, לפי התכונות של המ.אקספ, בה"ג, בכל איטרציה  $j$  נקבל ליטרל  $\ell_j$  עם ציון לפחות  $-\frac{1}{\varepsilon} \log(d)$  נסמן

$$\Delta = \frac{1}{\varepsilon} \log(d)$$

על כמה נקודות חיוביות נטעה בסך הכל?

אם הציון של ליטרל הוא לפחות  $-\Delta$ , אז בפרט הוא מתאפס על לכל היותר  $\Delta$  נקודות חיוביות. לכן בסה"כ אוסף כל  $J$  הליטרלים שנמצא טועים על לכל היותר  $J\Delta$  נקודות חיוביות.

על כמה נקודות שליליות נטעה?

נסתכל על איטרציה שבה  $|X^0| > 2k\Delta$ .

מכיוון שהציון של הליטרל הנבחר הוא לפחות  $-\Delta$ , אז אנחנו יודעים שהוא מתייג נכון (כלומר מתאפס על) לפחות

$$\frac{|X^0|}{k} - \Delta = \frac{|X^0| - k\Delta}{k} \geq \frac{|X^0| - |X^0|/2}{k} = \frac{|X^0|}{2k}$$

נקודות שליליות.

זכרו כי בתחילת כל איטרציה,  $|X^0|$  זהו מספר הנקודות השליליות "שעדיין לא תפסנו" כלומר מספר הנקודות השליליות שהליטרלים שיש לנו עד עכשיו טועים עליהם.

אז מה קיבלנו?

כל עוד מספר הנקודות השליליות שאנחנו טועים אליהן הוא לפחות  $2k\Delta$ , אז כל איטרציה מקטינה בפקטור לפחות  $\left(1 - \frac{1}{2k}\right)$  את מספר הנקודות השליליות שאנחנו טועים עליהן. לכן, עבור  $J$  מספיק גדול, מספר הנקודות שאנחנו טועים עליהן חייב לרדת מתחת ל  $2k\Delta$ .

איזה ערך  $J$  יספיק לנו? יספיק לנו לקחת  $J$  המקיים

$$n \cdot \left(1 - \frac{1}{2k}\right)^J \leq 1$$

מה שמתקיים אם

$$n \cdot e^{-J/2k} \leq 1$$

מה שקורה אם"

$$J \geq 2k \cdot \ln(n)$$

לסיכום, נבחר  $J \approx k \cdot \ln(n)$  ונקבל שבסיום הריצה ההשערה שאנחנו מחזירים טועה על לכל היותר

$$J\Delta \approx \frac{k^2 \ln^2(n)}{\varepsilon} \log(d)$$

נקודות חיוביות, וטועה על לכל היותר

$$2k\Delta \approx \frac{k^2 \ln(n)}{\varepsilon} \log(d)$$

נקודות שליליות.

לכן, אם נרצה שמספר הטעויות הזה יהווה לכל היותר חלק  $\alpha$  ממספר נקודות הקלט, אז נדרוש ש-

$$\underbrace{\alpha n}_{\substack{\text{מספר נקודות} \\ \text{שמותר לטעות עליהן}}} \gtrsim \underbrace{\frac{1}{\varepsilon} k^2 \ln^2(n) \log(d)}_{\substack{\text{חסם על מספר} \\ \text{הנקודות שטועים עליהן}}}$$

כלומר נדרוש

$$n \gtrsim \frac{1}{\alpha \varepsilon} k^2 \ln^2(n) \log(d)$$

קיבלנו אלגוריתם פרטי יעיל חישובית עם סיבוכיות מדגם משופרת. בפרט, עכשיו אנחנו צריכים דטה בגודל  $k^2$  במקום  $d^2$  (ואפשר לשפר ל-  $k^{1.5}$  עם קומפוזיציה חזקה...)

**הערה:** קיים אלגוריתם פרטי יעיל חישובית עם סיבוכיות מדגם (כמעט) כמו של המכניזם האקספוננציאלי.