הרצאה

היום נראה עוד משהו על מימ' גבוהים ועוד נעבור על ה-random sampling.

נדבר על **מעטפות תחתונות** ב-3D - $\mathbb{R}^3$

יש לנו $f_1,\dots,f_n$, $z=f_i(x,y)$

ואנו מעוניינים ב-$E(x,y)$, כרגיל. $E(x,y)=\min_i f_i(x,y)$

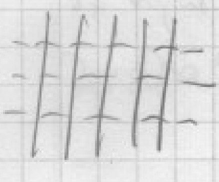
נגדיר את ה-minimization diagram, שזה פשוט הטלה של E על מישור xy. נקרא לזה M

נניח שיש לנו $F=\{f_1,\dots,f_n\}$
$G=\{g_1,\dots,g_n\}$
עם E_F, E_G, M_F, M_G.

במישור xy מקבלים overlay של M_F ו-M_G

ונשאלת השאלה, מה סיבוכיות ה-overlay? כמה קודקודים, צלעות פאות וכו'.
נתמקד בעיקר בקודקודים (ומהם ניתן להסיק על השאר).

ל-M_F ול-M_G יש סיבוכיות לכל היותר (כ"א) $O(n^{2+\varepsilon})$. ראינו זאת.



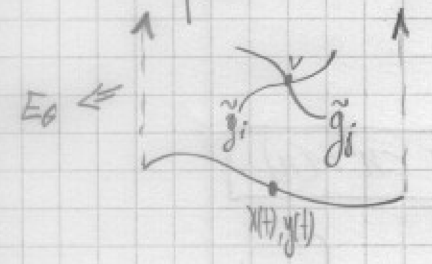
היינו מצפים באופן נאיבי שבשילוב משהו עם a צלעות עם משהו עם b צלעות שנקבל ab צלעות.
ולכן באופן נאיבי $O(n^{4+\varepsilon})$.

אבל באופן מפתיע, ל-overlay יש סיבוכיות $O(n^{2+\varepsilon})$ לכל ε.

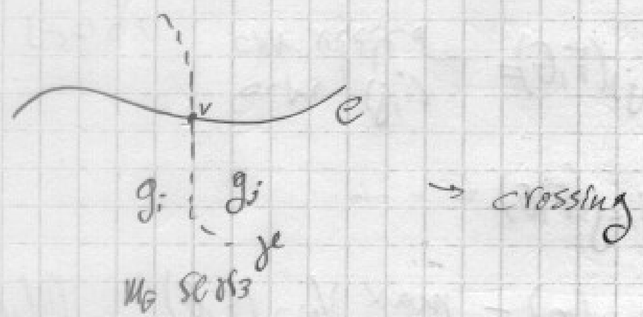
נרצה לפתח שוב נוסחת נסיגה. באופן דומה למה שעשינו שבוע שעבר.

ניקח צלע e (Fix an Edge) של M_F ונבנה וילון אנכי מעל e בכיוון z.

נסתכל על g_i על e: $\tilde{g}_i(t)=g_i(x(t),y(t))$



ואותו ציור במישור xy:

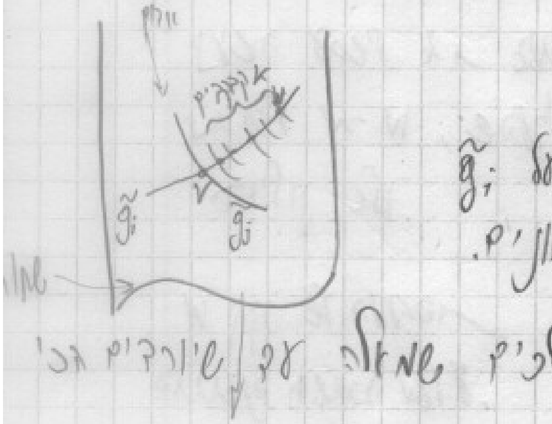


הסיבוכיות של E_G מעל e $\le O(\lambda_s(n))$.

אז אם יש לנו $n^{2+\varepsilon}$ קודקודים ולכן גם צלעות ומעל כל צלע כזו יכולים להיות מעבר מעטפת של $O(\lambda_s(n))$ אז כבר קיבלנו שסה"כ החיתוכים (crossings) בין הזוגות לשאריים: $O(n^{2+\varepsilon}\lambda_s(n)) = O(n^{3+\varepsilon})$

נשתמש ב- charging scheme בתוך V_e = הוילון של e.

שוב נניח ש- e מונוטוני בכיוון x, ונסתכל מה הולך ימינה ומה שמאלה.

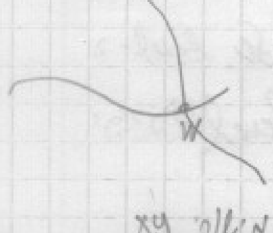


k פרמטר קבוע כלשהו.

נניח שמוצאים k קודקודים לא על המעטפת של $\tilde{g}_i$

charge-1 v ל- k הקודקודים הראשונים.

ה- charging הזה יחודי. לכל קודקוד הולכים שמאלה עד שיורדים הכי למטה (עד כדי קבוע של 2).



ה- level של w הוא לכל היותר k במערך האדום ו-0 במערך השחור.

זו ההטלה על מישור yz

נסתכל על החיתוכים בין צלע שחורה (חיתוך של 2 f-ים) ועם צלע אדומה (חיתוך של 2 g-ים. או להפך).

ה- level של v הוא (i,j) אם i פונק' שחורות ו- j פונק' אדומות נמצאות מתחת ל- v.

אנו רוצים את # הקודקודים ברמה $(0,0)$.

הגדרה: $V_{(i,j)}(F,G) =$ כמות הקדקדים ברמה (i,j)

$$V_{\le i,\le j}(F,G) = \dots$$

$$V_{(i,j)}(n) = \max V_{(i,j)}(F,G) \qquad (|F|,|G| \le n)$$

אבחנה (שיטת ה-charging):

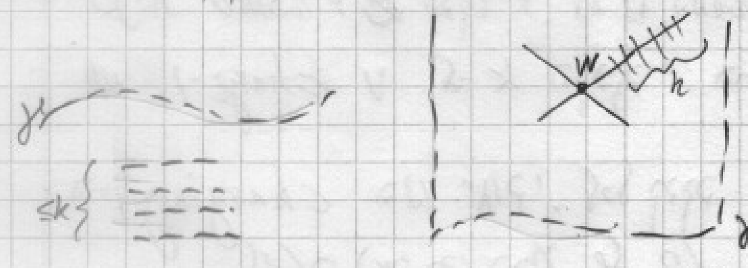
$V_{0,(\dots)}$?

$$V_{0,0}(n) \le \frac{2}{k} V_{0,\le k}(n) + \underbrace{\dots}_{\text{עוד סיפורים שהם במולטיפליקציה}}$$

האינטואיציה היא להשתמש בקלרקסון שוב, אבל אם נעשה את זה עכשיו, נקבל k^3, כי 4 פונק' מגדירות תצ"ה, ומחלקים ב-k. אנו רוצים לחלק ב-k^2.

$$V_{0,0} \le ck^2 V_{0,0}(n/k)$$

נעשה עכשיו את אותו דבר אבל נחליף בין אדום ושחור. כמו מתחיל מ-w, שהקשתות שלו יוצאות מהבלוק ונחפש קדקדים שאורים על הגליון שלו.



w קדקוד של המעטפת התחתונה השחורה מעל g.

במישור g

ה-level של w על השחור הוא 0. לכן לכל אלו שמסומנים גם יהיה black level לכל היותר k.

$$\Rightarrow V_{0,\le k}(n) \le \frac{2}{k} V_{\le k,\le k}(n) + \underbrace{\dots}_{\text{בבחירת ... הנים}}$$

$$\Rightarrow V_{0,0}(n) \le \frac{4}{k^2} V_{\le k,\le k}(n) + \dots$$

מעל משהו ברמה k אדומה ו-k שחורה יש לכל היותר $2k$ עקומות מתחתיו, והיות וכל crossings מוגדר ע"י 4 עקומות,

$$\Rightarrow V_0(n) \le \frac{4}{k^2} V_{\le 2k}(n) = \frac{4}{k^2} O\left(k^4 V_{00}\left(\frac{n}{k}\right)\right) \Rightarrow V_{0,0}(n) = O\left(k^2 V_{00}\left(\frac{n}{k}\right)\right)$$

ומזה מקבלים בסדר $n^{2+\varepsilon}$

בד"כ במדובר למימדים גבוהים, d זה מספק (לא עובד...)

דבר אחרון שמיכה רוצה לציין, בלי הוכחה.

Single Cell במערך של n משטחים ב-R^d.

עבור היפר-מישורים זה קל. d ישר יכול להופיע על היפר פעם אחת ויקבל פוליטופ, או פוליהדרא עם לכל היותר n facets - שוות d-1 מימדית.

הסיבוכיות הכוללת היא $O(n^{\lfloor d/2 \rfloor})$. ואז זה כמו למעטפת קמורה.

אבל, במקרה הכללי (לא מישורים) הסיבוכיות היא $O(n^{d-1+\varepsilon})$, כמו מעטפת תחתונה. הטופולוגיה והקומה הרבה יותר מורכבת עם כל מיני מסדרונות וכו'... ואפילו משולשים וסימפלקסים.

מכאן גם נובע ש-zone במקרה הכללי, הוא מסובכות $O(n^{d-1+\varepsilon})$. למה? מאותה סיבה כמו פעם - שמשטח יעבור ויתפוס ויגרום הרבה תאים לכל תא אחר. התוכנית - כמעט גבוה מסיבוכיות קבועה וכו'.

ולנושא אחר לגמרי.

Random Sampling

רוב התאוריה קמה ב-Machine Learning מתוך ג"ה השאלה על העניין.
המטרה: יש set עם המון עצמים ומקווים שע"י דגימה אקראית, נוכל לקבל מדגם המייצג די טוב את הסט המקורי.

לדוגמה: ישר עם n נק'. נדגום r מתוכם באקראי.

בוחרים כל נק' בדגימה באופן ב"ת בהסתברות $\frac{r}{n}$. בתוחלת זה r.

S = הסט המקורי של n הנק'.

R = דגימה.

באיזה מובן R מייצג טוב את S? ובכן, למשל אפשר להסתכל על אינטרוולים.

עבור אינטרוול I , נרשום $I_S = |S \cap I| / |S|$

$I_R = |R \cap I| / |R|$

ומה שמעניין בכלל זה היחס. שהיחסים יהיו קרובים.

הבעיה שבשביל קו, זה קטע overshoot, כי אם רוצים r נק' מייצגות פשוט ניקח כל נק' בהפרש $\frac{n}{r}$. ניקח את נק' $\frac{n\cdot 1}{r}, \frac{n\cdot 2}{r}, \dots, \frac{n\cdot r}{r}$ ונקבל הצגה מושלמת.

האינטואיציה אומרת, שעבור R לא קטן מדי, I_R ו-I_S יהיו קרובים, לכל I. לא ש-I ספציפי, אלא לכל I.

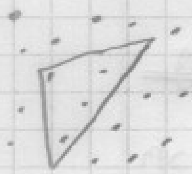
נגדיר Range Space - זוגות של (X, Γ).

נגדיר את X להיות ה-Ground Set של n עצמים. ישרים בעלי מימד, נק' על קו, וכו'-אבר.

Γ הוא סט של ranges, אובייקטים, כך שכל אחד הוא תת קבוצה (subset) של X.

למשל Range יכול להיות תחום על ישר ואז נקבל נק' 7 עד נק' 20.

או Range יכול להיות עבור נק' על מישור - משולש.



אבל לא כל subset הוא range.

נסתכל על דגימה $X \supseteq N$ בגודל r (דטרמיניסטי, או לא...)

אנו משתמשים במודל Samplingmodel כלומר באקראיות ובאופן בת"ת.

לכל range - $\Gamma \ni \varphi$ נגדיר:

$$\varphi_X = \frac{|X \cap \varphi|}{|X|}$$

למרות שברור ש-$\varphi = X \cap \varphi$ כי הרי φ הוא תת קבוצה של X. אבל חושבים על זה במובן הגיאומטרי, למשל ש-φ הוא משולש העוזר לנו לבחור את העצמים ב-X

$$\varphi_N = \frac{|N \cap \varphi|}{|N|}$$

<u>והמטרה</u>: להראות שלכל $\varepsilon>0$, לכל $\gamma\in\mathcal{R}$: $|\gamma_M-\gamma_x|<\varepsilon$

בהסתברות גבוהה. נסמן ב-r, כלומר ל-r גדול, ההסתב' תגדל)

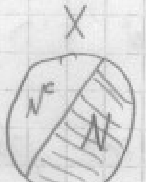
והשאלה הי' כמה גדול צריך להיות r, או ליתר דיוק, מה היחס בין r ל-ε?

תת קבוצה N המקיימת תכונה זו נקראת ε-קירוב (ε-approximation) של X (בהנתן ranges ב-$\mathcal{R}$).

<u>למשל</u>:

אם $\mathcal{R}=2^X$ - כלומר כל ה-subsets.

אפילו כש N של $\frac{n}{2}$ נק' לא יצליח

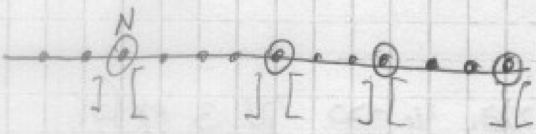


ועוד מושג: <u>ε-net</u>

N הוא ε-net אם $\gamma_x<\varepsilon \Leftarrow \gamma_N=0$. (המטרה)

$\equiv$ אם γ מכיל יותר מ-$\varepsilon|X|$ נק', אז הוא חייב להכיל נק' של N.

$\equiv$ N מנקב (פוגע) בכל ה-ε-heavy ranges.



אם נדגום r נק', מה יהיה ה- largest gap?

בממוצע כל הרווחים יהיו $\frac{n}{r}$ אבל בהסתברות גבוהה: $\text{largest gap} \le c\frac{n}{r}\log r$.

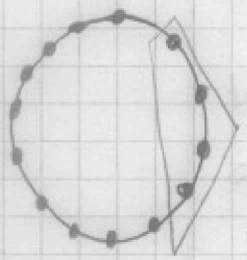
ואם נרצה שזה יהיה ε-net אז $\frac{cn}{r}\log r<\varepsilon n$ ← $\varepsilon\sim\frac{1}{r}\log r$ או $\boxed{r\sim\frac{c}{\varepsilon}\log\frac{1}{\varepsilon}}$

שזה יפה שמספר הנק' שיש לדגום הוא ב"ת ב-n.

כל הנושא הזה התחיל במאמר של VC - Vapnik & Chec- ב-1971. זה היה בכלל מאמר בסטטיסטיקה, והם הוכיחו תסם... הם דיברו על ε-קירוב.

לאחר מכן, Haussler & Welzl, ב-1987 דיברו על ε-nets.

וצריך איזושהי מגבלה על F. ראינו שאחרת, יכולים להיות מקרים שלא ניתן לעשות כלום.



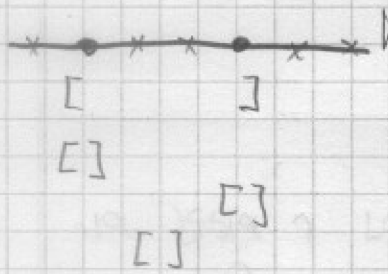
נסתכל לדוגמא על (X,F) כאשר, X הוא co של נק' על מעגל, ו-F co של כל ה-convex sets במישור.

אבל זה על הפנים כי לכל קב' של נק' על המעגל, נגדיר את הקמור שלהם ולכן הוא convex set והוא ב-F.
אז כאן אין תקווה לקבל ε-net טוב. לכל דגימה שניקח (אלא אם היא ממש הכל) נוכל לבחור co של נק' שלא כלול בו.

לכן, הגדירו את <u>VC-dimension</u>.
ראשית נגדיר: קב' A⊆X מנופצת (shattered), אם כל subset A'⊆A יכול להיווצר ע"י A∩R, עם R∈F כלשהו.

באופן הגדרה: ה- VC-dim = גודל הסט הכי גדול שיכול להיות מנופץ.

<u>דוגמא:</u>



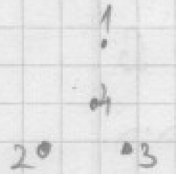
אבל עם 3 נק' כבר אי אפשר לנפץ כל תת קבוצה.

<u>עוד דוגמא:</u>

חצאי מישורים, ונק' במישור. 3 נק' זה לא בעיה. אבל אם נצייר

אפשר לנפץ כל תת קבוצה חוץ מ 14 ו- 23.
כי הם יכילו גם נק' נוספת. לא ניתן לייצר 1,4 ע"י חצאי מישורים.

ואם נצייר במצב לא קמור:



לא ניתן לנפץ את 4 ואת 123. אז VC-dim=3
וגם עם משולשים (במקום חצאי מישורים) ה- VC-dim=3

נשים לב ש- Vcdim יכול להיות סופי ואינסופי - אין גבול על
גודל גוף הקבוצה שניתן לנפץ.
כמו במצב הנק' של מעגל.

בד"כ נעסוק בבעיות עם vc-dim אינסופי. הם כאילו ליברליים מדי.
זה על הפנים.

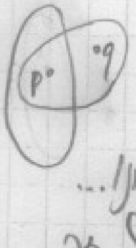
נסמן לבסוף את ה- vc-dim $\triangleq \delta$.

$\frac{n^\delta}{\delta!}$

ישנה למה חשובה: Sauer-Shelah
אם $n=|X|$, $\delta=$vc-dim, אז # ה-ranges הוא לכל היותר: $\binom{n}{0}+\binom{n}{1}+\dots\binom{n}{\delta}$
שזה בקירוב $O(n^\delta)$.
ולכן # ה-ranges הינו פולי' ב-n. בניגוד ל- 2^n שזה כמות כל
ה subsets.

ובכלל, זה ~~...~~ מה ששווה דברים. vc-dim הוא קצת מסובך וכו'. מה
שחשוב הוא שכמות ה-ranges היא פולי' ב-n.

אפשר למצוא את הוכחת הלמה באלף מקומות, כולל וויקיפדיה. הרעיון הוא באינדוקציה
על δ.



$\delta=1$ - לא ניתן לנפץ סטים בגודל 2.
נראה האם יש סט שמכיל רק את q... ואם לא,
אז לכל range שמכיל אותו, מכיל גם את p, אז ניתן להתעלם ממנו...
או שבכלל שמתחילים מ- $\delta=0$. בכל אופן כדאי להסתכל על זה.

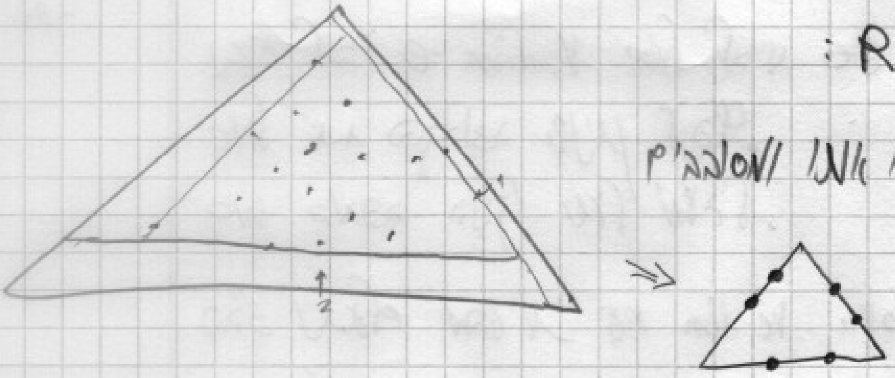
בשלב המעבר:
מסווגים כל range כאלו שמכילים את p (Σ_1) וכאלו שלא (Σ_2).

$$|\Sigma_2| \le \binom{n-1}{0}+\binom{n-1}{1}+\dots\binom{n-1}{\delta}$$

מישהו יסיים את זה פעם הבאה.

לשאול,

נק' ומשולשים ב- R^2:



לוקחים משולש מצמצמים אותו ומסובבים אותו

ולכן $O(n^6)$

אבל הוכחנו שה- VC-dim הוא 7.

ואפשר גם להסתכל על זה הפוך.

אם ה- corollary של הלמה נכון, כלומר יש n^δ טווחים, אז ניתן לומר משהו על ה- VC-dim.

אז אם A, בגודל x הוא הכי גדול שניתן לנפץ, אז # ה- ranges ב- A הוא 2^x והרי $2^x \le x^\delta$ $\Leftarrow$ $x \le \delta \log x$ $\rightarrow$ $x = O(\delta \log \delta)$

VC הראו: אם דוגמים $\frac{c\delta}{\varepsilon^2}\log\frac{1}{\varepsilon}$ נק', אנו מקבלים ε-קירוב בהסתב' גבוהה.

HW הראו: אם דוגמים $\frac{c\delta}{\varepsilon}\log\frac{1}{\varepsilon}$ נק', אנו מקבלים ε-net בהסתב' גבוהה.