הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ד – 3 בספטמבר 2024

מרצה: ד"ר ניר רוזנפלד

מבחן מסכם מועד א'

הנחיות הבחינה:

- **משך הבחינה:** שלוש שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מחשבון: מותר.
- כלי כתיבה: עט בלבד.
- יש לכתוב את התשובות **על גבי שאלון זה**.
- מותר לענות בעברית או באנגלית.
- הוכחות והפרכות צריכות להיות פורמליות.
- קריאות:
  - סימונים לא ברורים בשאלות רב-ברירה ו/או תשובות מילוליות בכתב יד לא קריא יובילו לפסילת התשובה.
  - לא יתקבלו ערעורים בנושא.
- במבחן 19 עמודים ממוספרים סה"כ, כולל עמוד שער זה שמספרו 1 ושלושה עמודי טיוטה בסוף הגיליון.
- נא לכתוב רק את המבוקש ולצרף הסברים קצרים עפ"י ההנחיות.
- **בתום המבחן יש להגיש את שאלון זה בלבד.**
- **לזכאים להערכה חלופית מתאפשרת בחירה בין שאלות 3 ו-4.**
- **זכרו: Less is more**. אל תכתבו פרטים מיותרים.

בהצלחה!

שאלה 1: SVM, Imbalanced sampling [35 נק']

בבעיה זאת נעסוק בבעיית סיווג בינארי $\mathcal{Y}=\{-1,1\}$, עם מרחב דוגמאות $\mathcal{X}=\mathbb{R}^d$.

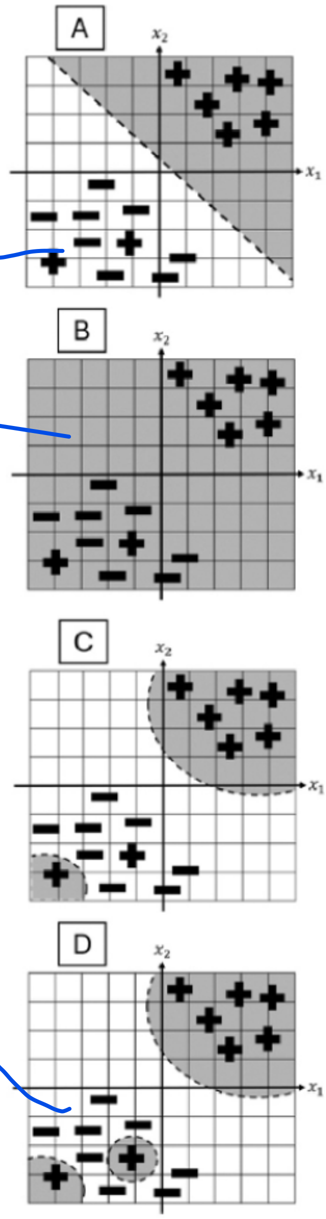
א. [15 נק׳] בסעיף זה $d=2$. דגמו באקראי $S\subset\mathcal{X}$.
מימין נתונים 4 איורים המראים decision boundaries שונים. באזורים אפורים המודל מנבא $+1$.
משמאל נתונות 4 גרסאות שונות של בעיות SVM. בנוסף נתונה פונקציה למיפוי פיצ׳רים פולינומיאלית ממימד גבוה ϕ.

1) ע״י מתיחת קו, לכל בעיית SVM התאימו איור אחד שמייצג את ה-decision boundary של המודל שיתקבל מבעיה זו. אם בעיית ה-SVM לא יכולה להגיע לפתרון, סמנו עליה [X].

תזכורת: כלל ההחלטה עבור מפריד לינארי ללא מיפוי ϕ - $h(\mathbf{x})=sign(\mathbf{w}^\top\mathbf{x}+b)$.

עם מיפוי ϕ - $h(\mathbf{x})=sign(\mathbf{w}^\top\phi(\mathbf{x})+b)$.

I) $\underset{\mathbf{w},b}{\operatorname{argmin}}\ \\|\mathbf{w}\\|_2^2+b$
II) $\underset{\mathbf{w},b}{\operatorname{argmin}}\ \\|\mathbf{w}\\|_2^2$ such that: $y_i(\mathbf{w}^\top\mathbf{x}_i+b)\geq 1,\forall i$
III) $\underset{\mathbf{w},b}{\operatorname{argmin}}\ \\|\mathbf{w}\\|_2^2$ such that: $y_i(\mathbf{w}^\top\phi(\mathbf{x}_i)+b)\geq 1,\forall i$
IV) $\underset{\mathbf{w},b}{\operatorname{argmin}}\ \\|\mathbf{w}\\|_2^2+\sum_i \max\{0,1-y_i(\mathbf{w}^\top\mathbf{x}_i+b)\}$



2) נמקו בקצרה את בחירתכם.

I) 0 vector
II)no graph
III) hard svm fof high degree
IV) soft SVM

3) מבין ארבעת בעיות ה- SVM, עבור מי הביטוי $\mathbb{E}_{S,x}\left[\left(h_S(x)-\bar{h}(x)\right)^2\right]$ הוא הגדול ביותר? נמקו בקצרה.

הערה: התוחלת כאן היא ביחס למשתנים המקריים S, x.

תזכורת: הפונקציה $\bar{h}(x)$ היא המסווג הממוצע של מחלקת ההיפותזות המוגדרת ע״י בעיית ה-SVM.

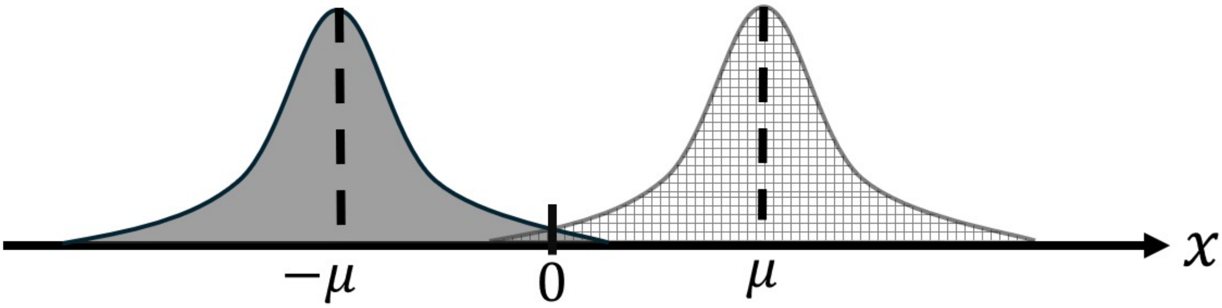
III) highest complexity -> highest Variance

ב. [20 נק׳] בסעיפים הבאים ננתח תרחיש של imbalanced sampling, תרחיש שבו התפלגות התיוגים בקבוצת מדגם S היא לא מאוזנת.

ננתח את המקרה שבו $d = 1$. ידוע כי הפיצ׳ר x מתפלג כמו משתנה אקראי גאוסי, כאשר תוחלת הגאוסיאן תלויה בערך של y. השונות זהה בשני המקרים. יהי $\mu > 0$, אזי:

- $\mathbb{P}(x|y=1) = N(\mu, \sigma^2)$.
- $\mathbb{P}(x|y=-1) = N(-\mu, \sigma^2)$.
- $\mathbb{P}(y=1) = \mathbb{P}(y=-1) = 1/2$.

1) להלן שרטוט של התפלגות המשותפת (עם צביעה שונה להתפלגויות המותנות השונות):



נניח כי יש לנו גישה להתפלגות עצמה, כלומר אנחנו יודעים את μ, σ^2. מהו המפריד מסוג threshold (לינארי) אשר ישיג את שגיאת ההכללה המינימאלית על פילוג זה.

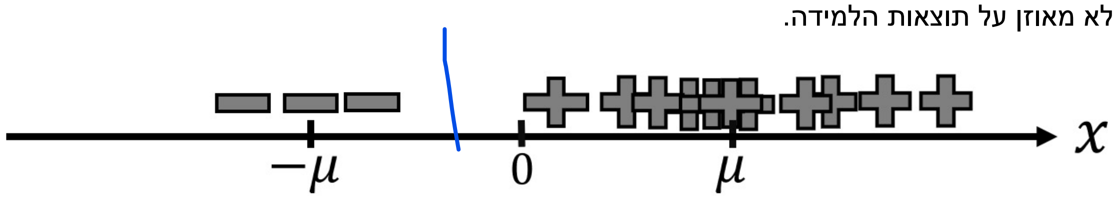
כתבו את כלל ההחלטה שלו (כלומר כלל הפרדיקציה $\hat{y} = h(x)$) באופן מפורש, ונמקו בקצרה.

h(x) = sign(x)

נניח כעת כי יש לנו גישה רק לדגימה מההתפלגות (כלומר אנחנו לא יודעים את $(\mu,\sigma^2$). יהי $S=\{(x_i,y_i)\}_{i=1}^{n}$ מדגם בגודל n אשר כל דגימה (x_i,y_i) נדגמה $i.i.d$ מההתפלגות המשותפת $\mathbb{P}(x,y)$. נסמן ב-n_+ את מספר הדוגמאות החיוביות במדגם, וב-n_- את מספר הדוגמאות השליליות, כך ש- $n=n_++n_-$.

נגיד שהמדגם "לא מאוזן" כאשר n_+ גדול משמעותית מ-n_- (או להפך).

2) להלן איור של מדגם מסוים S שאינו מאוזן. נניח שלמדנו בעזרת מסווג threshold עם אלגוריתם SVM. סמנו על האיור את המקום בו עובר המפריד הנלמד. נמקו את בחירתכם, והסבירו את ההשלכות של מדגם לא מאוזן על תוצאות הלמידה.



will be right in the middle point of the right most (-) and the leftmost (+) as a result of trying to maximize the margin

as a result it will have a higher generalization error and will even be mistaken on (-) that are between the \theta and 0

3) עבור התפלגות נתונה ומדגם בגודל n ממנה, נגדיר את הגדלים הבאים:

- $M(n)=\mathbb{E}[\max(x_1,\dots,x_n)]$.
- $m(n)=\mathbb{E}[\min(x_1,\dots,x_n)]$.

כאשר מדובר בהתפלגות גאוסיאנית עם תוחלת μ (ושונות קבועה σ^2), נסמן ב-$M(n;\mu)$ ו-$m(n;\mu)$ את הגדלים המתאימים.

נתון שמתקיים בקירוב: $M(n;0)\cong\sigma\sqrt{2\ln(n)}$, $m(n;0)\cong-\sigma\sqrt{2\ln(n)}$.

השלימו:

$M(n;\mu)\cong$ M(n;0) + \myu	$M(n;-\mu)\cong$ M(n;0) - \myu
$m(n;\mu)\cong$ m(n;0) + \myu	$m(n;-\mu)\cong$ m(n;0) - \myu

4) נתון שדגמתם מדגם לא מאוזן אך פריד. שיטה נפוצה להתמודדות עם בעיה זו נקראת subsampling.
במסגרת שיטה זו, מוציאים באקראי מ-S דוגמאות **בעלות התיוג הנפוץ ביותר**, עד שמקבלים קבוצת מדגם מאוזנת , כלומר עד ש-$n_+ = n_-$.

השתמשו בסעיפים הקודמים, ובפרט בסעיף (3), על מנת להסביר כיצד subsampling יכול לסייע לאימון SVM בתרחיש של שאלה זו.
הדרכה: שימו לב כי הגדלים $M(n), m(n)$ תלויים ב- n. בנוסף זכרו מה התכונה של המפריד הלינארי שמחזיר SVM.

this will lower the amount of samples until the Excpected Max & Min values will be of the same (seder godel)
therefore the Expected (mafrid) will be better generalized
and we'll get the optimal one (sign(x)) as a result of (M(n;0)-\myu + m(n;0)+\myu)\2 = 0;

שאלה 2: VC-dimension [25 נק']

א. [2 נק'] להלן ההגדרה של "ניתוץ". השמטנו מההגדרה את הַכַּמָּתִים.

השלימו את שלושת הכמתים החסרים. בכל מקום כתבו בבירור האם חסר בהגדרה ∀ או ∃.

$$\mathcal{H} \text{ shatters } C \iff \underbrace{\text{A}}_{\text{השלימו}} y_1, \dots, y_{|C|} \in \mathcal{Y}: \underbrace{\text{E}}_{\text{השלימו}} h \in \mathcal{H}: \underbrace{\text{A}}_{\text{השלימו}} x_i \in C: h(x_i) = y_i$$

בשאלה זו נתון מרחב דוגמאות $\mathcal{X}$ כלשהו ומרחב תיוגים $\mathcal{Y} = \{-1, +1\}$. אם נדרשת הוכחה, הוכיחו באופן פורמלי. אם נדרשת הפרכה, תנו דוגמה נגדית מנומקת היטב והוכיחו כי היא אכן מפריכה את הטענה.

ב. [5 נק'] הוכיחו/הפריכו.

יהיו שתי מחלקות $\mathcal{H}_1, \mathcal{H}_2$. אזי

$$VCdim(\mathcal{H}_1 \cup \mathcal{H}_2) \geq \max\{VCdim(\mathcal{H}_1), VCdim(\mathcal{H}_2)\}$$

proof:
let max{VC(H1),VC(H2)} = T
let's say VC(H1) >= VC(H2);
therefore there is a set of T points x1,...,xt that for each possible labelling y1,...yt there exists a h in H1 s.t h(xi) = yi
let's look at T and H1vH2; H1 shatters T - therfore all hypothesises h that help H1 shatter T are also in H1vH2; therefore H1vH2 shatters T and VC(H1vH2) >= T = max(....,...)

ג. [6 נק'] הוכיחו.

נתונה מחלקה $\mathcal{H}$ סופית. אזי

$$VCdim(|H|) \leq \log_2 |H|.$$

הוכחה פורמלית:

let VC(H) = d
then there exists a set of x1,...,xd st for all y1,...,yd there exists h in H st forall i: h(xi) = yi
therefore H >= 2^d => logH >= d = VC(H)

ד. [12 נק'] הוכיחו/הפריכו.

יהיו $\mathcal{H}_1, \mathcal{H}_2$ מחלקות ונתונה $\emptyset \neq C \subset \mathcal{X}$. נניח

1) $\mathcal{H}_1$ מנתצת את C.

2) $\mathcal{H}_2$ אינה מנתצת את C.

אזי בהכרח

$$VCdim(\mathcal{H}_1) > VCdim(\mathcal{H}_2)$$

הוכחה פורמלית:

H1 = {h(x1) = 1 ^ h(x2) = -1, h'(x1) = -1 ^ h'(x2) = 1}

H2 = {h''(x1) = 1 ^ h''(x2) = 1, h'''(x1) = 1 ^ h'''(x2) = -1}

VCdim(H1) <= logH1 = 1, let's take x = x1; H1 shatters it => VC(H1) = 1
same for H2 except x = x2; VC(H2) = 1;

let C be {x1}, H1 shatters it while H2 doesn't, yet VC(H1) = VC(H2)
therefore we disproved the claim

שאלה 3: Perceptron [20 נק']

☐ לזכאים להערכה חלופית בלבד (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של יתר השאלות יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

בשאלה זו נניח $\mathcal{Y}=\{+1,-1\}$ ומרחב דוגמאות $\mathcal{X}=\mathbb{R}^d$. לפניכם אלגוריתם הפרספטרון כפי שהוא הוצג בתרגול.

```
input: training set S = {(x_1, y_1), ..., (x_m, y_m)},  step size η = 1

w = 0_d
while did not separate the training set:
      for i = 1 to m:
            ŷ_i = sign(w^T x_i)
            if y_i ≠ ŷ_i:
                  w = w + y_i x_i
return w
```

לפניכם **משפט** התכנסות הפרספטרון:

תהי קבוצת אימון $S=\{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_m,y_m)\}$. נניח כי קיימים וקטור משקולות w_* כך ש- $\|\mathbf{w}_*\|_2=1$ ו- $\gamma>0$ כך שלכל $i=1,\dots,m$ מתקיים,

$$y_i(\mathbf{w}_*^\top\mathbf{x}_i)\ge\gamma$$

נניח בנוסף כי לכל $i=1,\dots,m$ מתקיים $\|\mathbf{x}_i\|_2\le R$.

אזי אלגוריתם הפרספטרון עושה לכל היותר $\frac{R^2}{\gamma^2}$ טעויות.

בשאלה זו נוכיח משפט זה בשלבים. תהי S קבוצת אימון ונניח את קיומם של γ,R,w_* כמו במשפט.

נגדיר את $\mathbf{w}_k$ להיות וקטור המשקולות שנלמד עד הטעות ה- k (לא כולל).

שימו לב: עבור הגדרה זו מתקיים $\mathbf{w}_1=\mathbf{0}_d$.

א. [2 נק'] האם S פרידה לינארית? נמקו בקצרה.

yes because there is a gamma > 0 st : y(wx) >= gamma > 0 therefore it (mesaveg) each point correctly therefore it is seperable

ב. [5 נק׳] נתחיל בלהוכיח באינדוקציה את הטענה הבאה:

$$\mathbf{w}_{k+1}^{\top}\mathbf{w}_{*} \geq k\gamma$$

<u>הדרכה:</u>

- בדקו את בסיס האינדוקציה עבור $k = 0$.
- הניחו את נכונות הטענה עבור k כלשהו, כלומר: $\mathbf{w}_{k}^{\top}\mathbf{w}_{*} \geq (k-1)\gamma$.
- הוכיחו באמצעות כלל העדכון של הפרספטרון כי $\mathbf{w}_{k+1}^{\top}\mathbf{w}_{*} \geq \mathbf{w}_{k}^{\top}\mathbf{w}_{*} + \gamma$.
- השלימו את צעד האינדוקציה.

w_(k+1) * w* = (w_(k) + yx)w* >= (k-1)g + g = kg as wanted

ג. [2 נק׳] לרשותכם אי-שוויון קושי שוורץ

$$\|\mathbf{u}\|_2\|\mathbf{v}\|_2 \geq |\mathbf{u}^{\top}\mathbf{v}| \qquad \mathbf{u} \in \mathbb{R}^d, \mathbf{v} \in \mathbb{R}^d$$

הסיקו

$$\|\mathbf{w}_{k+1}\|_2 \geq \mathbf{w}_{k+1}^{T}\mathbf{w}_{*}$$

סה״כ מהסעיפים הקודמים מתקבל $\|\mathbf{w}_{k+1}\|_2 \geq k\gamma$. כעת ניגש לחסם מלעיל של הביטוי.

ד. [9 נק׳] הוכיחו באינדוקציה את הטענה

$$\|\mathbf{w}_{k+1}\|_2^2 \leq kR^2$$

<u>הדרכה:</u>

- בדקו את בסיס האינדוקציה עבור $k = 0$.
- הניחו את נכונות הטענה עבור k כלשהו, כלומר: $\|\mathbf{w}_k\|_2^2 \leq (k-1)R^2$.
- הוכיחו באמצעות כלל העדכון של הפרספטרון כי $\|\mathbf{w}_{k+1}\|_2^2 \leq \|\mathbf{w}_k\|_2^2 + R^2$.
- השלימו את צעד האינדוקציה.

<u>זכרו:</u> עבור $\mathbf{v} \in \mathbb{R}^d$ מתקיים $\|\mathbf{v}\|_2^2 = \mathbf{v}^\top\mathbf{v}$.

סה״כ מכל הסעיפים עד כה קיבלנו

$$k^2\gamma^2 \leq \|\mathbf{w}_{k+1}\|_2^2 \leq kR^2$$

ה. [2 נק׳] הסיקו כעת את המשפט. כלומר:

$$k \leq \frac{R^2}{\gamma^2}$$

כאשר k הוא מספר הטעויות.

שאלה 4: Deep Learning [20 נק']

☐ לזכאים להערכה חלופית <u>בלבד</u> (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של יתר השאלות יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

בשאלה זו נעבוד מעל מרחב דוגמאות $\mathcal{X}=\mathbb{R}$ ומרחב תיוגים $\mathcal{Y}=\{0,1\}$. בסעיפים בהם נדרש חישוב, עגלו את תשובותיכם לדיוק של עד שלוש ספרות אחרי הנקודה.

נתונה רשת נוירונים שעוצבה לפתירת בעיית סיווג בינארי המוגדרת ע״י הקשרים הבאים:

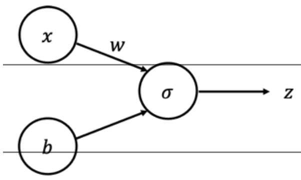
$$\begin{bmatrix} a_1 = & w_{11}x+b_{11} \\ a_2 = & w_{12}x+b_{12} \\ z_1 = & ReLU(a_1) \\ z_2 = & ReLU(a_2) \\ a_3 = & w_{21}z_1+w_{22}z_2+b_{21} \\ \hat{y} = & \sigma(a_3) \end{bmatrix}$$

כאשר $x\in\mathbb{R}$ הוא הקלט ו- $\hat{y}\in[0,1]$ הוא הפלט של הרשת.

האימון של הרשת מתבצע באמצעות cross-entropy loss. תזכורת:

$$\begin{bmatrix} ReLU(x) = & \max(0,x) \\ \sigma(x) = & \dfrac{1}{1+e^{-x}} \\ \ell^{CE}(y,\hat{y}) = & -ylog(\hat{y})-(1-y)\log(1-\hat{y}) \end{bmatrix}$$

א. [2 נק'] ציירו תרשים של הרשת (לרשותכם דוגמה לתרשים בגוף התשובה).
ציינו מי הן המשקולות של הרשת שביחס אליהן אנו עושים אופטימיזציה.



$\boxed{z=\sigma(xw+b)}$

ב. [5 נק'] נניח כי

$$\begin{bmatrix} b_{11} = & 0.04 \\ b_{12} = & 0.01 \\ b_{21} = & 0.08 \\ w_{11} = & 0.2 \\ w_{12} = & -0.1 \\ w_{21} = & 0.7 \\ w_{22} = & 0.7 \end{bmatrix}$$

עבור קלט $x = 0.3$ חשבו את $\hat{y}$. אם ידוע כי הרשת מנבאת 1 אמ"מ $\hat{y} > \frac{1}{2}$, מה היא תנבא עבור $x = 0.3$?

ג. [1 נק׳] באיזה כלל אנו משתמשים בשביל לחשב את הנגזרות החלקיות ביחס למשקולות של רשת נוירונים כלשהי? סמנו את התשובה הנכונה:

1. כלל הפיצה.
2. כלל האצבע.
3. כלל השרשרת.
4. כלל השורש.

בסעיף הבא הניחו כי עבור הדוגמה $x = 0.3$ התיוג האמיתי שלה הוא $y = 1$.

נעסוק כעת באלגוריתם ה- backpropogation.

ד. [12 נק'] בצעו את אלגוריתם backpropogation על המשתנה b_{12}. עליכם לכתוב את הנגזרת החלקית של פונקציית ההפסד $\ell(y, \hat{y})$ ביחס למשתנה b_{12}, דהיינו $\frac{\partial \ell}{\partial b_{12}}$, באמצעות הנגזרות החלקיות $\frac{\partial \alpha}{\partial \beta}$, כאשר α, β יכולים להיות כל אחד מהבאים:

$$\ell, \hat{y}, z_i, a_i, b_{ij}, w_{ij}, x$$

עבור כל הערכים החוקיים של i, j. וודאו כי כל נגזרת חלקית $\frac{\partial \alpha}{\partial \beta}$ לא ניתנת לפירוק לנגזרות חלקיות פשוטות יותר.

לאחר מכן, **חשבו** את $\frac{\partial \ell}{\partial b_{12}}$ עבור $x = 0.3$. <u>תזכורת</u>: $\frac{d}{dx}\sigma(x) = \sigma(x) \cdot \left(1 - \sigma(x)\right)$

$\frac{\partial \ell}{\partial b_{12}} =$

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):