פרספטרון - יוצר על-מישור שמפריד דוגמאות במרחב תוך הנחה שהדוגמאות בקבוצת האימון **ניתנות להפרדה בצורה לינארית.** הפרספטרון מוצג בתור $h(\mathbf{x}) = sign\left(\sum_{i=1}^{d} w_i x_i + b\right)$ עם $w_0 = b$ ו- $x_0 = 1$ נקבל $sign\left(\mathbf{w}^T\mathbf{x}\right)$. נאתחל את $\hat{\mathbf{w}}(0) = 0$ (איטרציה 0) נחפש (אקראית/סדרתית) נק' שלא תויגה נכון - משמע $y(t)\hat{\mathbf{w}}^T(t)\mathbf{x}(t) < 0$ ונעדכן ע"פ $\hat{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) + y(t)\cdot\mathbf{x}(t)$ כתוצאה מכך נקבל את הסיווג החדש של הנקודה $y(t)\mathbf{x}^T(t)\hat{\mathbf{w}}(t+1) = y(t)\mathbf{x}^T(t)\cdot[\hat{\mathbf{w}}(t) + y(t)\cdot\mathbf{x}(t)] = y(t)\mathbf{x}^T(t)\hat{\mathbf{w}}(t) + y^2(t)\|\mathbf{x}(t)\|^2$ ומכך ש- $x_0 = 1$ אז $\|\mathbf{x}(t)\|^2 > 0$ מה שאומר שהסיווג החדש של הנקודה שלנו יהיה **גדול** יותר מהסיווג הקודם (לא מבטיח שיפור כללי, אלא רק עבור הנקודה): $y(t)\mathbf{x}^T(t)\hat{\mathbf{w}}(t+1) > y(t)\mathbf{x}^T(t)\hat{\mathbf{w}}(t)$ כך שלבסוף (כתוצאה מההנחה על פרידות המידע) נקבל $\rho = \min y_n \cdot \left(\mathbf{w}^{*T}\mathbf{x}_n\right) > 0$ כתוצאה מכך האלגוריתם מתכנס. הגורמים להתכנסות ארוכה - (1) הדוגמה עם הנורמה המקסימלית (2) המרווח הקטן ביותר בין 2 דוגמאות שמתויגות שונה. עבור **המקרה בו לא ניתן להפריד לינארית** נוכל להגביל את מס' האיטרציות ולשמור את המפריד הטוב ביותר מבין האיטרציות עם אלגוריתם pocket - שבו בכל איטרציה מעדכנים את המשקלים, מחשבים $E_{in} = \frac{1}{N}\sum_{n=1}^{N}[\![sign\left(\mathbf{w}^T\mathbf{x}_n\right) \neq y_n]\!]$ ועל פיו שומרים את המשקלים של המפריד הטוב ביותר. **במקרה בו המידע פריד אך לא לינארית** ניתן לבצע טרנס' לא-לינארית לפיצ'רים $\mathbf{z} = \Phi(\mathbf{x})$ למרחב $\mathcal{Z}$ בו המידע פריד לינארית. אפשרי לאפטם את PLA על פי $E_{in} = \frac{1}{N}\sum_{n=1}^{N}\max(0, -y_n\mathbf{w}^T\mathbf{x}_n)$ בעזרת SGD. המפריד שיבחר תלוי ב: 1: אתחול $\mathbf{w}$, 2: סדר הצגת הדוגמאות.

רגרסיה לינארית - רוצים להביא למינימום את פונקציית המחיר $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{\mathbf{w}}^T\mathbf{x}_n\right)^2$. מגדירים את המטריצה $X_{N\times(d+1)}$ ששורותיה הן וקטורי הפיצ'רים (התצפיות) והעמודה הראשונה היא עמודת אחדות ואת וקטור התיוגים $\mathbf{y}_{N\times 1}$, ואז: $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\|\mathbf{y} - X\hat{\mathbf{w}}\|^2 =$ $(*)$ $\frac{1}{N}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^TX\hat{\mathbf{w}} + \hat{\mathbf{w}}^TX^TX\hat{\mathbf{w}}\right]$ נגזור ונשווה ל-0: $\nabla_{\hat{\mathbf{w}}}E_{in} = \frac{1}{N}\left[2X^TX\hat{\mathbf{w}} - 2X^T\mathbf{y}\right] = \frac{2}{N}X^T\left(X\hat{\mathbf{w}} - \mathbf{y}\right) = 0$ נשמיט את הקבוע ונקבל את **המשוואות הנרומליות**: **(1)** $X^TX\hat{\mathbf{w}} = X^T\mathbf{y}$ ו- **(2)** $X^T\left(X\hat{\mathbf{w}} - \mathbf{y}\right) = 0$. (נרצה לקרב את $X\hat{\mathbf{w}}$ ל- $\mathbf{y}$ כך שהשגיאה $(X\hat{\mathbf{w}} - \mathbf{y})$ **תהיה אורתוג' לעמודות** X ול-$X\hat{\mathbf{w}}= \hat{\mathbf{y}}$). נוכל לגזור שוב ולקבל $\nabla^2_{\hat{\mathbf{w}}}E_{in} = \frac{2}{N}X^TX \succeq 0$ כי $\mathbf{a}^TX^TX\mathbf{a} = (X\mathbf{a})^T(X\mathbf{a}) = \|X\mathbf{a}\|^2 \geq 0$ ולכן **קונבקסית וכל נקודה שתקיים את המשוואות הנורמליות היא מינ' גלובלי.** בהנחה ש X^TX הפיכה (X דרגה מלאה) הפתרון למשוואות הנורמליות נובע מהכפלה משמאל בהפכית לה: $\left(X^TX\right)^{-1}X^TX\hat{\mathbf{w}} = \left(X^TX\right)^{-1}X^T\mathbf{y}$ נקבל פתרון סגור: $\hat{\mathbf{w}} = \left(X^TX\right)^{-1}X^T\mathbf{y}$ ונבחין כי כיוון שהפיכה $X^TX\mathbf{c} \neq 0$ ואז $\mathbf{c}^TX^TX\mathbf{c} \neq 0$ ולכן $X^TX \succ 0$ - מכך נסיק שהיא **קונבקסית ממש** והפתרון שמצאנו הוא מינ' גלובלי **יחיד!** החזאי יוגדר בתור: $\hat{\mathbf{y}} = X\hat{\mathbf{w}} = X\cdot\left(X^TX\right)^{-1}X^T\mathbf{y}$ כאשר $H = X\left(X^TX\right)^{-1}X^T$ היא **מטריצה המטילה את $\mathbf{y}$** (סימטרית ו- $H^2 = H$) **על עמודות** X **כש-** $\hat{\mathbf{y}}$ ההיטל. השגיאה הסופית שמתבססת על המשוואות הנורמליות תניב לנו את משפט פיתגורס: $\|\hat{\mathbf{y}} - \mathbf{y}\|^2 =$ $\|\mathbf{y} - X\hat{\mathbf{w}}\|^2 = (\mathbf{y} - X\hat{\mathbf{w}})^T(\mathbf{y} - X\hat{\mathbf{w}}) \overset{(2)}{=} \mathbf{y}^T(\mathbf{y} - X\hat{\mathbf{w}}) = \|\mathbf{y}\|^2 - \mathbf{y}^TX\hat{\mathbf{w}} \overset{(1)}{=} \|\mathbf{y}\|^2 - \hat{\mathbf{w}}^TX^TX\hat{\mathbf{w}} = \|\mathbf{y}\|^2 - \|\hat{\mathbf{y}}\|^2$ ואם X לא מדרגה מלאה אז X^TX לא הפיכה ויש אינסוף פתרונות עבור $\hat{\mathbf{w}}$ שההפרש ביניהם בגרעין של X. אך פתרון יחיד עבור $\hat{\mathbf{y}} = X\hat{\mathbf{w}}$ כי הוא היטל ולכן כל הפתרונות של $\hat{\mathbf{w}}$ משיגים את אותה שגיאה מינימלית כי E_{in} קונבקסית ו $X^TX \succeq 0$ ויש נק' מינ'. דרך נוספת לפתרון היא **השלמה ריבועית** עבור $(*)$ כאשר מניחים ש- $X^TX \succ 0$: לביטוי $\hat{\mathbf{w}}^TX^TX\hat{\mathbf{w}}$ קל להגיע אך צריך להוסיף את $-2\mathbf{y}^TX\hat{\mathbf{w}}$ - נתחיל מ- $(\hat{\mathbf{w}} - \mathbf{b})^TX^TX(\hat{\mathbf{w}} - \mathbf{b})$ ועל מנת שיתקיים: $-\mathbf{b}^TX^TX\hat{\mathbf{w}} = -\mathbf{y}^TX\hat{\mathbf{w}}$ נגדיר: $\mathbf{b} = \left(X^TX\right)^{-1}X^T\mathbf{y}$, מציבים, מוסיפים קבועים ומקבלים את הביטוי הריבועי: $\frac{1}{N}\left[(\hat{\mathbf{w}} - \left(X^TX\right)^{-1}X^T\mathbf{y})^T(X^TX)(\hat{\mathbf{w}} - (X^TX)^{-1}X^T\mathbf{y}) + \|\mathbf{y}\|^2 - \mathbf{y}^TX(X^TX)^{-1}X^T\mathbf{y}\right]$ כאשר $X^TX \succ 0$ והכפל השמאלי בסוגריים המרובעים אי-שלילי והבאתו למינימום לפי $\hat{\mathbf{w}}$ מחייבת שיתאפס משמע $\mathbf{c}X^TX\mathbf{c} = \mathbf{o}$ אם"ם $\mathbf{c} = 0$ ולכן: $\hat{\mathbf{w}} = \left(X^TX\right)X^T\mathbf{y}$ ואז: $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\left[\mathbf{y}^T\left(I - X\left(X^TX\right)^{-1}X^T\right)\mathbf{y}\right]$ וגם $I - X\left(X^TX\right)^{-1}X^T$ מטריצת הטלה. אם נסמן $X = \begin{bmatrix}\mathbf{X}_0 & \ldots & \mathbf{X}_d\end{bmatrix}$ (עמודה ראשונה היא אחדות) ונסמן $\hat{\mathbf{y}} = X\hat{\mathbf{w}}$ אז מ-(1) נקבל ש- $X^T\hat{\mathbf{y}} = X^T\mathbf{y}$ בפרט מתקיים שוויון בעמודה הראשונה: $\mathbf{1}^T\hat{\mathbf{y}} = \mathbf{1}^T\mathbf{y}$ מה שמשליך על כך **שהממוצעים של ערכי $\hat{\mathbf{y}}$ ו- $\mathbf{y}$ שווים**: $\frac{1}{N}\sum_{n=1}^{N}\hat{y}_n = \frac{1}{N}\sum_{n=1}^{N}y_n$ **וממוצע השגיאות הוא 0**. **נקודה חשובה:** אם היינו יודעים את התוחלת $E\left[\left(y - \mathbf{x}^T\hat{\mathbf{w}}\right)^2\right]$ היינו יודעים את פונקציית **צפיפות הפילוג המשותפת** ואז היינו מקבלים את הפתרון האופטימלי על ידי גזירה: $\frac{\partial}{\partial\hat{\mathbf{w}}}E[\cdot] = E\left[-2\mathbf{x}\left(y - \mathbf{x}^T\hat{\mathbf{w}}\right)\right] = -2\left(E[\mathbf{x}y] - E\left[\mathbf{x}\mathbf{x}^T\right]\hat{\mathbf{w}}\right) = 0 \Rightarrow \hat{\mathbf{w}} = \left(E\left[\mathbf{x}\mathbf{x}^T\right]\right)^{-1}\cdot E[\mathbf{x}y]$ כאשר $E[\mathbf{x}y]$ היא הקורלציה בין $\mathbf{x}$ ל- y ו- $E\left[\mathbf{x}\mathbf{x}^T\right]$ של $\mathbf{x}$ עם עצמו. והמשערך האופטימלי מושתת על **התוחלת הסטטיסטית**, אך מכיוון שאנחנו לא יודעים את פונ' הצפיפות - אנחנו מנסים להביא למינימום את ממוצע שגיאות המרחקים שהוא מושתת במשערך האופטימלי על **תוחלת אמפירית** (כל התוחלות במשערך האופטימלי הופכות לממוצעים). אפשרי לבצע **טרנס' לא לינארית** על מנת לעבור למרחב אחר $\mathcal{Z}$ בו המטריצת שורות שלנו: $X = x_i^j, i = 1,\ldots,N,\ j = 0,\ldots,d$ (עמודה ראשונה 1) היא מטריצת ונדרמונד שדרגתה כמס' ה-x_i-ים השונים זה מזה.

רגרסיה לוגיסטית - רוצים להניב $\hat{y} \in [0,1]$ ע"י הלבשת סיגמויד $\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$ על המשערך כאשר: $(1)\ 1 - \theta(s) = \theta(-s)$ ו- $(2)\ \theta'(s) = \theta(s)\cdot(1 - \theta(s))$ - משמע $h(\mathbf{x}) = \theta\left(\mathbf{w}^T\mathbf{x}\right)$ ורוצים ללמוד את ההסתברות $P(y = 1 \mid \mathbf{x})$. משתמשים בשגיאת cross - entropy שלוקחת 2 פונ' הסתברות של משתנים בינארים: $(p, 1-p), (q, 1-q)$ ומודדת את המרחק ביניהן. פונקציית השגיאה $\left(h(\mathbf{x}_n) := \theta\left(\mathbf{w}^T\mathbf{x}_n\right)\right)$: $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N}I(y_n = +1)\log\left(\frac{1}{\theta(\mathbf{w}^T\mathbf{x}_n)}\right) + I(y_n = -1)\cdot\log\left(\frac{1}{\theta(-\mathbf{w}^T\mathbf{x}_n)}\right)$ - עבור $y_n = +1$ נרצה $h(\mathbf{x}_n) = \theta\left(\mathbf{w}^T\mathbf{x}_n\right) \to 1$ ועבור $y_n = -1$ נרצה $1 - h(\mathbf{x}_n) = \theta\left(-\mathbf{w}^T\mathbf{x}_n\right) \to 1$. עם זאת נבחין שאפשרי להכניס את ערכי y_n לתוך θ (התוצאה לא תשתנה) ונקבל: $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N}I(y_n = +1)\log\left(\frac{1}{\theta(y_n\mathbf{w}^T\mathbf{x}_n)}\right) + I(y_n = -1)\cdot\log\left(\frac{1}{\theta(y_n\mathbf{w}^T\mathbf{x}_n)}\right)$ ואם נוציא את log החוצה נקבל: $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{\theta(y_n\mathbf{w}^T\mathbf{x}_n)}\right)[I(y_n = +1) + I(y_n = -1)]$ סכום האנדיקטורים הוא 1 ונקבל את פונקציית השגיאה הסופית של cross-entropy: $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N}\log\left(1 + e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}\right)$ אותה נרצה להביא למינימום, נגזור: $\frac{\partial}{\partial\hat{\mathbf{w}}}E_{in}(\hat{\mathbf{w}}) = \frac{\partial}{\partial\hat{\mathbf{w}}}\frac{1}{N}\sum_{n=1}^{N}\log(1 + e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}) = \frac{1}{N}\sum_{n=1}^{N}(1 + e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n})^{-1}\cdot e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}\cdot(-y_n\mathbf{x}_n)$ נייפה: $\nabla_{\hat{\mathbf{w}}}E_{in}(\hat{\mathbf{w}}) = -\frac{1}{N}\sum_{n=1}^{N}\frac{e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}}{1+e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}}\cdot(y_n\mathbf{x}_n)$ ונראה שאין פתרון סגור: $-\frac{1}{N}\sum_{n=1}^{N}\theta\left(-y_n\hat{\mathbf{w}}^T\mathbf{x}_n\right)(y_n\mathbf{x}_n) = 0$ ולכן פונים ל Gradient - Descent. נבחין גם כי זו פונקציה קונבקסית כי הנגזרת השנייה (הסיאן): $\nabla^2E_{in} = \frac{1}{N}\sum_{n=1}^{N}\theta\left(-y_n\hat{\mathbf{w}}^T\mathbf{x}_n\right)\left(1 - \theta\left(-y_n\hat{\mathbf{w}}^T\mathbf{x}_n\right)\right)(-y_n\mathbf{x}_n)^2$ אי-שלילית ולכן GD תמצא מינ' גלובלי. אפשרי להשתמש ברגרסיה

לוגיסטית לסיווג והיא טובה מפרספטרון כי יש לה מינ' גלובלי ויהיה קל לאפטם אותה וגם אין צורך בכך שהמידע יהיה פריד. $(**)$ ניתן גם למדל את הבעיה בצורה שונה: נסמן $h(\mathbf{x}_n) = p(y_n = 1 \mid \mathbf{x}_n)$ ואז בעזרת הרכבת exp על log יחס ההסתברויות נגיע לסיגמויד: $\log\left(\frac{h(\mathbf{x}_n)}{1-h(\mathbf{x}_n)}\right) = \mathbf{w}^T\mathbf{x}_n \xrightarrow{\exp()} h(\mathbf{x}_n) = (1 - h(\mathbf{x}_n))\cdot e^{\mathbf{w}^T\mathbf{x}_n} \Rightarrow h(\mathbf{x}_n)\left(1 + e^{\mathbf{w}^T\mathbf{x}_n}\right) = e^{\mathbf{w}^T\mathbf{x}_n}$ ולכן $h(\mathbf{x}_n) = \theta$.

Gradient Descent - אלגוריתם איטרטיבי למציאת מינימום - בכל שלב נתבונן ב- $\Delta E_{in} = E_{in}(\hat{\mathbf{w}}(t+1)) - E_{in}(\hat{\mathbf{w}}(t))$ (הפרש בין השגיאה באיטרציה t ואיטרציה $t+1$). המשקלים באיטרציה $t+1$ מושפע מגודל צעד **קטן** η_t ווקטור יחידה $\hat{\mathbf{v}}_t$: $\hat{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) + \eta_t\hat{\mathbf{v}}_t$. נקבל: $\Delta E_{in} = E_{in}(\hat{\mathbf{w}}(t) + \eta_t\hat{\mathbf{v}}_t) - E_{in}(\hat{\mathbf{w}}(t))$ ונוכל להשתמש **בקירוב לינארי** עם **טור טיילור מסדר ראשון סביב** $\hat{\mathbf{w}}(t)$ בצורה הבאה: $\Delta E_{in} \approx E_{in}(\hat{\mathbf{w}}(t)) + \eta_t\nabla^TE_{in}(\hat{\mathbf{w}}(t))\hat{\mathbf{v}}_t - E_{in}(\hat{\mathbf{w}}(t)) = \eta_t\cdot\nabla^TE_{in}(\hat{\mathbf{w}}(t))\hat{\mathbf{v}}_t$ ועל מנת למזער את השגיאה **נבחר את** $\hat{\mathbf{v}}_t$ כוקטור יחידה נגד כיוון הגרדיאנט: $\Delta E_{in} \approx \eta_t\cdot\nabla E_{in}(\hat{\mathbf{w}}(t))\hat{\mathbf{v}}_t \geq \eta_t\cdot\nabla^TE_{in}(\hat{\mathbf{w}}(t))\cdot\frac{-\nabla^TE_{in}(\hat{\mathbf{w}}(t))}{\|\nabla^TE_{in}(\hat{\mathbf{w}}(t))\|} = -\eta_t\cdot\|\nabla E_{in}(\hat{\mathbf{w}}(t))\|$ כך המכפלה הפנימית תהיה מינימלית. **את η_t נבחר** בצורה פרופורציונלית לנורמה של הגרדיאנט (כי הנורמה קטנה ככל שמתקרבים למינ'): $\eta_t = \eta\cdot\|\nabla E_{in}(\hat{\mathbf{w}}(t))\|$ בתוספת בחירת $\hat{\mathbf{v}}_t$: $\hat{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) - \eta\cdot\nabla E_{in}(\hat{\mathbf{w}}(t))$. את האלגוריתם ניתן לאתחל ב- $\hat{\mathbf{w}}(0) = 0$ או באתחול רנדומלי כך שכל איבר בוקטור הוא הגרלה של משתנה מקרי נורמלי עם תוחלת אפס ושונות מאוד קטנה. **מבחינת סיום האלגו':** (1) מגדירים מס' איטרציות מקס' (2) מגדירים חסם ε שנורמת גרדיאנט נמוכה ממנו מספיקה לנו (3) אופציונלי - להוסיף חסם עליון לערך השגיאה. batch GD הוא מצב בו מחשבים את השגיאה על **כל הדוגמאות** ועל פיה מבצעים עדכון. Stochastic/online/sequential GD הוא מצב בו בוחרים דוגמה אחת (אפשר גם כמה) באופן אקראי (הסתברות אחידה) ומחשבים עבורה את השגיאה והגרדיאנט ועל פיו מעדכנים את המשקלים. בממוצע נבצע דבר דומה ל-batch כיוון שאנחנו בוחרים נקודות באופן אקראי אחיד והתוחלת תהיה זהה אך התהליך **מהיר פי** n! **ובורחים מנקודות מינימום מקומי.** Variable learning rate GD זה שימוש בהיריסטיקה שמשנה את קצב הלמידה η - מאתחלים את η_0 ומשקלים $\mathbf{w}(0)$. בכל איטרציה בודקים האם השגיאה קטנה, אם כן: מגדירים $\eta_{t+1} = \alpha\eta_t$ עבור $\alpha \approx 1.05 - 1.1$ (מגדילים את η), ומקבלים את עדכון המשקולות, אם לא: מגדירים $\eta_{t+1} = \beta\eta_t$ עבור $\beta \approx 0.5 - 0.8$ (מקטינים את η) ולא מקבלים את עדכון המשקולות, כך משתנים גדלי הצעדים שלנו בצורה דינמית. Steepest Descent זו שיטה בה בוחרים מראש η^* יחיד ואופטימלי על ידי הבאה למינימום של $\arg\min_\eta E_{in}(\mathbf{w}(t) + \eta\mathbf{v}(t))$ בכל איטרציה בה עוד לא הגענו לתנאי העצירה (לדוגמה מס' איטרציות).

SVM - אלגוריתם רובסטי, מובטחת הכללה. באלגוריתם זה משתמשים ב-bias ולכן מפרידים אותו מהמשקולות: $h(\mathbf{x}) = sign(\mathbf{w}^T\mathbf{x} + b)$. **בהנחה שהמידע פריד לינארית** נרצה על-מישור $\mathbf{w}^T\mathbf{x} + b = 0$ שמפריד **בצורה מושלמת** את הדוגמאות משמע $sign(\mathbf{w}^T\mathbf{x}_n + b) = y_n$ לכל n, ובצורה שקולה: $y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n + b\right) > 0$ לכל n. חישוב שולי העל-מישור: כל נקודה על העל-מישור היא כזו שתקיים את המשוואה שלו $\mathbf{w}^T\mathbf{x} + b = 0$ ועבור כל 2 **נק' (לא וקטורים)** $\mathbf{x}'$ ו- $\mathbf{x}''$ על העל-מישור L: $\mathbf{w}^T\mathbf{x}' + b = 0$ ו- $\mathbf{w}^T\mathbf{x}'' + b = 0$, נחסר את 2 המשוואות ונקבל $\mathbf{w}^T(\mathbf{x}' - \mathbf{x}'') = 0$ ולכן $(\mathbf{x}' - \mathbf{x}'')$ אורתוג' ל- $\mathbf{w}$. משמע $\mathbf{w}$ אורתוג' לכל **וקטור (הפרש בין 2 נק' על L)** מ- L ול-L עצמו. כעת נרצה לחשב את המרחק בין הנק' x ל-L (האנך), נבחר נק' $x' \in L$ ונחשב את גובה ההיטל של הוקטור $(\mathbf{x} - \mathbf{x}')$ על $\mathbf{w}$ (שמאונך ל-L) תוך נרמול $\mathbf{w}$ כך שנקבל את המרחק הרצוי: $d(\mathbf{x}, L) = \left|\frac{\mathbf{w}}{\|\mathbf{w}\|}(\mathbf{x} - \mathbf{x}')\right| = \frac{1}{\|\mathbf{w}\|}\cdot\left|\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{x}'\right| \overset{\mathbf{w}^T\mathbf{x}'=-b}{=} \frac{1}{\|\mathbf{w}\|}\cdot\left|\mathbf{w}^T\mathbf{x} + b\right|$. נרצה למצוא את העל-מישור המפריד **שימקסם את המרחק המינימלי** מהנקודות: $\max_{\mathbf{w},b}\left(\min_n\frac{1}{\|\mathbf{w}\|}\cdot\left|\mathbf{w}^T\mathbf{x}_n + b\right|\right)\ \mathbf{s.t}\ y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0(*)$ (השוליים יהיו סימטריים ביחס לדוגמאות הקרובות ביותר מכל תיוג). שוליים רחבים ביותר: בהינתן על-מישור $\mathbf{w}^T\mathbf{x} + b$ שמפריד בצורה מושלמת: (i) אם נגדיר $\rho = \min_n y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0$ נקבל: $\min_n y_n\cdot(\mathbf{w}^T\mathbf{x}_n + b)/\rho = 1$ ומכך שחלוקת על-מישור בערך קבוע לא משנה אותו נוכל להפוך את אילוץ האי-שוויון $y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0$ לאילוץ שוויון: $\min_n y_n\cdot(\mathbf{w}^T\mathbf{x}_n + b) = 1$. (ii) מכך ש- $y_n \in \{-1,1\}$ ומההנחה שמדובר בהפרדה מושלמת: $\left|\mathbf{w}^T\mathbf{x} + b\right| = \left|y_n\cdot(\mathbf{w}^T\mathbf{x}_n + b)\right| = y_n\cdot(\mathbf{w}^T\mathbf{x}_n + b)$. (iii) נציב את $(i) + (ii)$ ב-$(*)$ ונקבל שהבעיה: $\max_{\mathbf{w},b}\frac{1}{\|\mathbf{w}\|}$ שקולה לפתרון: $\min_{\mathbf{w},b}\|\mathbf{w}\| \equiv \min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2$ ומכך שהנחנו שהדוגמאות פרידות ושיש לפחות דוגמה אחת מכל תיוג נפשט את האילוץ $\min_n y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$ אותו אפשר להפוך ל- $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$, נראה שלא יכול להתקיים אי-שוויון חזק: נניח בשלילה שלכל n: $y_n(\mathbf{w}^{*T}\mathbf{x}_n + b^*) > 1$ ונגדיר $\rho^* = \min_n y_n(\mathbf{w}^{*T}\mathbf{x}_n + b^*) > 1$ ונגדיר $\mathbf{w} \leftarrow \mathbf{w}^*/\rho^*$ ו- $b \leftarrow b^*/\rho^*$ ונקבל: $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$ (לפחות אחת הדוגמאות מקיימת שוויון) בנוסף לכך ש- $\mathbf{w}^* \neq 0$ (כי ישנה הפרדה ולפחות 2 דוגמאות) נקבל: $\|\mathbf{w}\| = \frac{1}{\rho^*}\|\mathbf{w}^*\| < \|\mathbf{w}^*\|$ כי ρ^* גדול מ-1 אבל זו **סתירה** לכך ש- $\mathbf{w}^*$ הוא פתרון אופטימלי ולכן בהכרח $y_n\cdot(\mathbf{w}^{*T}\mathbf{x}_n + b^*) \geq 1$ ויש לפחות דוגמה אחת שמקיימת שוויון - לכן **בעית האופטימיזציה** של **Hard-Margin SVM** היא: $\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2\ \mathbf{s.t}\ y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1\ (1)$.
ניתן להגדיר את הבעיה כבעית Convex QP שהפתרון הוא $\mathbf{u}^* \leftarrow \text{QP}(Q, \mathbf{p}, A, \mathbf{c})$. נוכל נגדיר את: $\mathbf{p} = \mathbf{0}$, $\mathbf{u} = \begin{bmatrix} b & \mathbf{w}^T\end{bmatrix}^T$, $\mathbf{c} = \mathbf{1}$ ואת המטריצות $Q_{(d+1)\times(d+1)} = diag(0,1,1,\ldots,1)$ ו- $A = (diag(y_1,\ldots y_N)\cdot X)$ ששורותיה מהצורה $\mathbf{a}_n^T = y_n\cdot[1, \mathbf{x}_n^T]$ וזוהי בעית QP קונבקסית כי $\boldsymbol{\alpha}^TQ\boldsymbol{a} \geq 0$. נוכל לקבל $\mathbf{u}^* = [b^*, \mathbf{w}^*]$ ולקבל את החזאי האופטימלי $g(\mathbf{x}) = sign(\mathbf{w}^{*T}\mathbf{x} + b^*)$ בו כל נקודה על השול מקיימת את האילוץ של (1) בשוויון ומרחקה מהמפריד הוא $1/\|\mathbf{w}\|$. הבעיה הדואלית: נגדיר את $f_0 = \frac{1}{2}\mathbf{w}^T\mathbf{w}$ שהיא קונבקסית וכל הפונקציות שלנו אפיניות מהצורה $y_n\left(\mathbf{w}^T\mathbf{x} + b\right) \geq 1$ ומההנחה שהמידע פריד לינארית בהכרח יש פתרון ולכן **מתנאי ה-KKT לבעיה קונבקסית** נוכל להסיק דואליות חזקה. הלגרנג'יאן: $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N}\alpha_n\left(1 - y_n(\mathbf{w}^T\mathbf{x}_n + b)\right)$. גזירה לפי b: $\frac{\partial\mathcal{L}}{\partial b} = -\sum_{n=1}^{N}\alpha_ny_n = 0$, גזירה לפי $\mathbf{w}$: $\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N}\alpha_ny_n\mathbf{x}_n \Rightarrow \mathbf{w} = \sum_{n=1}^{N}\alpha_ny_n\mathbf{x}_n$. נציב את מה שקיבלנו ב- $\mathcal{L}$ ונקבל פונקציה אשר תלויה ב- $\boldsymbol{\alpha}$ בלבד: $g(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_my_ny_m\mathbf{x}_n^T\mathbf{x}_m + \sum_{n=1}^{N}\alpha_n - \sum_{n=1}^{N}\alpha_ny_n(\mathbf{w}^T\mathbf{x}_n + b)$ כאשר מהביטוי האחרון אפשרי להוציא את $\mathbf{w}$ ואת b החוצה ולקבל: $g(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_my_ny_m\mathbf{x}_n^T\mathbf{x}_m + \sum_{n=1}^{N}\alpha_n - \mathbf{w}\sum_{n=1}^{N}\alpha_ny_n\mathbf{x}_n - b\sum_{n=1}^{N}\alpha_ny_n$ וכאמור $\mathbf{w} = \sum_{n=1}^{N}\alpha_ny_n\mathbf{x}_n$ ו: $-b\sum_{n=1}^{N}\alpha_ny_n = 0$ בגלל גזירת b, נציב ונקבל: $g(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_my_ny_m\mathbf{x}_n^T\mathbf{x}_m + \sum_{n=1}^{N}\alpha_n$. ולכן **הבעיה הדואלית של ה- Hard Margin SVM**: $(2) max_{\boldsymbol{\alpha}} - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_my_ny_m\mathbf{x}_n^T\mathbf{x}_m + \sum_{n=1}^{N}\alpha_n\ \text{s.t}\ \boldsymbol{\alpha} \geq \mathbf{0},\ \boldsymbol{\alpha}^T\mathbf{y} = 0$ (לאחר החלפת סימן). נוכל להגדיר את הבעיה הדואלית כבעית **convex QP** מהצורה: $\min\frac{1}{2}\boldsymbol{\alpha}^TQ_D\boldsymbol{\alpha} + \mathbf{p}_D^T\boldsymbol{\alpha}\ \text{s.t}\ A_D\boldsymbol{\alpha} \geq \mathbf{c}$ (נשים לב שהחלפנו את האילוץ $\boldsymbol{\alpha}^T\mathbf{y} = 0$ בשני אילוצים $\boldsymbol{\alpha}^T\mathbf{y} \geq 0$ ו- $-\boldsymbol{\alpha}^T\mathbf{y} \geq 0$ וכך נוכל לשלב את שני האילוצים המקוריים באחד יחיד) נגדיר את Q_D: על פי אבריה: $Q_{Di,j} = y_iy_j\mathbf{x}_i^T\mathbf{x}_j$ והיא P.S.D כי אם נגדיר את מטריצה M: $M = diag(y_i\mathbf{x}_i)$ אז $Q_D = M^TM$ ולכן P.S.D, $\mathbf{c} = \mathbf{o}$,

$\mathbf{p}_D = \underline{-1}$, ו- $A_D = \begin{bmatrix} \mathbf{y}^T \\ -\mathbf{y}^T \\ I_{N\times N} \end{bmatrix}$. נקבל $\boldsymbol{\alpha}^* \leftarrow QP(Q_D, \mathbf{p}_D, A_D, \mathbf{c})$ שעל פיו מוגדר $\mathbf{w}^*$: $\mathbf{w}^* = \sum_{n=1}^{N} \alpha_n^* y_n \mathbf{x}_n$. מ-complementary slackness נקבל שכל אחת מהמכפלות ב- $\sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^T\mathbf{x}_n + b)\right)$ **שווה לאפס** ולכן אם עבור s כלשהו: $\alpha_s > 0$ אז האילוץ מתקיים בשוויון: $y_s(\mathbf{w}^T\mathbf{x}_s + b) = 1$, משמע **הנקודה היא על השוליים** והיא support vector. אם נציב בשוויון הנ"ל את $\mathbf{w}^*$ נקבל את ערך b^*: $b^* = y_s - \mathbf{w}^{*T}\mathbf{x}_s$ ואת החזאי שלנו: $g(\mathbf{x}) = sign(\mathbf{w}^{*T}\mathbf{x} + b^*)$ או בצורה מפורשת: $g(\mathbf{x}) = \sum_{\alpha_s^*>0} \alpha_s^* y_s \mathbf{x}_s^T \mathbf{x} + b^*$ **נקודות חשובות**: $\underline{1}$: אם $\alpha_i > 0$ אז $\mathbf{x}_i$ בהכרח על השוליים אך ההיפך לא נכון - אם $\mathbf{x}_i$ על השוליים לא בהכרח ש $\alpha_i > 0$ $\underline{2}$: אם $\mathbf{x}_i$ לא על השוליים אז $\alpha_i = 0$ $\underline{3}$: אם $\alpha_i = 0$ לכל i אז $\mathbf{w} = \mathbf{0}$ (חזאי קבוע) אבל מההנחה שיש לפחות 2 נקודות בעלות תיוג שונה וההפרדה מושלמת - זה לא אפשרי! **חשוב**: אם $\boldsymbol{\alpha}^*$ מאפטם את הבעיה הדואלית אז $\|\mathbf{w}^*\|^2 = \sum_{n=1}^{N} \alpha_n^*$ (מתקבל מדואליות חזקה שα^* מתאים גם לבעיה הפרימלית) והמרחק המינימלי בין נקודה למפריד האופטימלי: $(\sqrt{\sum_{n=1}^{N} \alpha_n^*})^{-1} = \frac{1}{\|\mathbf{w}\|}$. את **בעית האופטימיזציה** של **Soft-Margin SVM** נגדיר: $(3)\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} \xi_n \text{ s.t } y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1 - \xi_n, \xi_n \geq 0$ (בעית **Convex QP**) כאשר לכל נק' $\mathbf{x}_n$ ישנה ξ_n (slack variable) שמגדירה את ה"חדירה" של $\mathbf{x}_n$ אל תוך השוליים, משמע אם $\xi_n \in (0,1)$ אז הנקודה $\mathbf{x}_n$ מסווגת נכון אך **חצתה את השוליים**. אם $\xi_n > 1$ אז הנקודה $\mathbf{x}_n$ **לא מסווגת נכון**.(**) **מדוע** $\xi_n \not< 0$? כי אם כן אז נקודות אשר נמצאות תחת סיווג נכון "יחפו" על נקודות שנכנסו לשוליים ואז לא היינו נענשים עליהן. משמע **אם המידע לא פריד לינארית ו-**C **לא גדול מידי - תמיד נוכל למצוא פתרון פיסיבילי.** עבור $C \to \infty$ ייבחר על-מישור בעל שוליים מאוד צרים בעל פחות הפרות של השוליים **ועבור מידע פריד ו-**C כלשהו מספיק גדול **הפתרון של Soft-Margin יתלכד עם הפתרון של Hard-Margin**. עבור $C \to 0$ יבחר על-מישור עם שוליים מאוד רחבים (אין הענשה) ויותר הפרות של השוליים. נפתח את הלגרנג'יאן (אחרי הבאת האילוצים לצורה הנכונה): $\mathcal{L}(\mathbf{w}, v, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} \xi_n + \sum_{n=1}^{N} \alpha_n\left(1 - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)\right) - \sum_{n=1}^{N} \beta_n\xi_n$. נגזור את $\mathcal{L}$ על פי כל ξ_i: $\frac{\partial\mathcal{L}}{\partial\xi_i} = C - \alpha_i - \beta_i \Rightarrow \alpha_i + \beta_i = C$ ועבור b: $\frac{\partial\mathcal{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0$ ועבור $\mathbf{w}$ נקבל את: $\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \Rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$. נציב חזרה בלגרנג'יאן: ונראה שלאחר צמצום נקבל את אותו ביטוי כמו במקרה של Hard-Margin SVM: $-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n\alpha_m y_n y_m \mathbf{x}_n^T\mathbf{x}_m + \sum_{n=1}^{N} \alpha_n$ **והבעיה הדואלית של Soft-Margin SVM** היא: $\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n\alpha_m y_n y_m \mathbf{x}_n^T\mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \text{ s.t } \boldsymbol{\alpha}_n \geq 0,\ \boldsymbol{\beta}_n \geq 0,\ \alpha_n + \beta_n = C,\ \sum_{n=1}^{N} \alpha_n y_n = 0$ ולאחר סידור האילוצים: $(4)max_{\boldsymbol{\alpha},\boldsymbol{\beta}} - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n\alpha_m y_n y_m \mathbf{x}_n^T\mathbf{x}_m + \sum_{n=1}^{N} \alpha_n \text{ s.t } \boldsymbol{\alpha}_n \geq 0,\ \boldsymbol{\alpha}_n \leq C,\ \sum_{n=1}^{N} \alpha_n y_n = 0$ זוהי בעיה ששקולה לבעיה הדואלית של ה-Hard למעט האילוץ $\alpha_n \leq C$ ניתן לייצגה כבעית **Convex QP** ולקבל פתרון $\boldsymbol{\alpha}^* \leftarrow QP(Q_D, \mathbf{p}_D, A_D, \mathbf{c}_D)$ שעל פיו $\mathbf{w}^* = \sum_{a_s^*>0} \alpha_s^* y_s \mathbf{x}_s$ וה comp. slackness שלנו: $\alpha_n(1 - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)) = 0$ ו- $\beta_n \cdot \xi_n = (C - \alpha_n)\xi_n = 0$ לכל $n = 1, \ldots, N$. כאשר $\underline{\alpha_s^* > 0}$ אז: $y_s(\mathbf{w}^{*T}\mathbf{x}_s + b^*) = 1 - \xi_s \leq 1$ (משמע שהנקודה $\mathbf{x}_s$ על השוליים, בתוך השוליים או שחצינו את המפריד). כאשר $\underline{\beta_s > 0}$ בעצם מדובר ב- $\underline{\alpha_s^* < C}$ ואז: $\xi_s = 0$ וכתוצאה מכך $y_s(\mathbf{w}^{*T}\mathbf{x}_s + b^*) \geq 1$ **ולכן**: אם נשלב את שני התנאים: $\underline{0 < \alpha_s^* < C}$ אז בהכרח $y_s(\mathbf{w}^{*T}\mathbf{x}_s + b^*) = 1$ והנקודה $\mathbf{x}_s, y_s$ היא על השוליים (margin support vectors). אפשר למצוא גם כאן את b^* (נזכור ש: $y_s^2 = 1$): $y_s(\mathbf{w}^{*T}\mathbf{x}_s + b^*) = 1 \iff y_s^2(\mathbf{w}^{*T}\mathbf{x}_s + b^*) = y_s \iff b^* = y_s - \mathbf{w}^{*T}\mathbf{x}_s$. אם $\underline{\alpha_s^* = C}$ אז $\mathbf{x}_s$ על השוליים או בתוכם ונקראת bounded support vector. **נקודה חשובה**: לא מחייב שיהיה קיים $\alpha_s^* \in (0, C)$ שכן במקרה בו סט הדוגמאות: $\mathcal{D} = \{((2,3), 1), ((2,2), -1), ((1,2), 1), ((1,3), -1)\}$ ועבור $C = 10$ נקבל שאף נקודה לא תקיים $\alpha_s^* \in (0, C)$ וכמו-כן $\mathbf{w} = \mathbf{0}$, $\boldsymbol{\alpha} = (10, 10, 10, 10)$, $b = -1$ ו- $\boldsymbol{\xi} = (0, 2, 0, 2)$ מהווים פתרונות פיסיביליים (פרימלי ודואלי). עבור מידע **פריד אך לא לינארית** נוכל להגדיר טרנס' לא ליני' $\phi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$, להפעיל אותה על כל נקודה $\mathbf{x}_n \in \mathbb{R}^d$ ולקבל נק' חדשה $\phi(\mathbf{x}_n) = \mathbf{z}_n \in \mathbb{R}^{\tilde{d}}$ ואז לפתור את הבעיה $\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} \text{ s.t } y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$ עבור $\mathbf{w} \in \mathbb{R}^{\tilde{d}}$, ע"י ניסוחה כבעית **Convex QP** ולקבל פתרון $\mathbf{w}^*, b^*$ ואת החזאי: $g(\mathbf{x}) = sign(\mathbf{w}^{*T}\phi(\mathbf{x}) + b^*)$ במרחב של $\mathbf{z}_n$. הפתרון יוגדר כראוי: המפריד יהיה בעל שוליים רחבים, המרחק בין הנקודות על השוליים לבין המפריד יהיה $1/\|\mathbf{w}\|$ וההפרדה תהיה מושלמת - **אמנם** כשנחזור לממד של $\mathbf{x}_n$ המרחקים ישתנו וה support vectors יהיו במרחקים שונים מהמפריד אך הצלחנו להשיג מפריד לא לינארי בעל שוליים. כמו-כן כיוון שמדובר ב-SVM, השוליים שלו עוזרים ביצירת **הכללה** טובה, מה שיצמצם התאמת-יתר גם בממדים גבוהים! **בעיה** קיימת במעבר עם ϕ היא העובדה שבמעבר לממדים גדולים מאוד בעית ה-QP קשה מאוד מבחינה חישובית ולכן נוכל לעבור **לבעיה הדואלית** שם אנחנו תלויים **בכמות הדוגמאות ולא במימדן** ($\boldsymbol{\alpha} \in \mathbb{R}^N$) ונקבל: $\min_{\tilde{\alpha}} \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\alpha}_n\tilde{\alpha}_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N} \tilde{\alpha}_n \text{ s.t } \tilde{\alpha}_n \geq 0, \sum_{n=1}^{N} \tilde{\alpha}_n y_n = 0$ והפתרון יניב לנו $\tilde{\mathbf{w}}^* = \sum_{n=1}^{N} \tilde{\alpha}_n y_n \mathbf{z}_n$ ואת $\tilde{b}^*$ ונקבל את החזאי $g(\mathbf{x}) = sign(\tilde{\mathbf{w}}^{*T}\phi(\mathbf{x}_n) + \tilde{b}^*) = sign(\sum_{n=1}^{N} \tilde{\alpha}_n y_n \phi(\mathbf{x}_n)^T\phi(\mathbf{x}_n) + \tilde{b}^*)$. עם זאת, עדיין צריך לחשב את $\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_n)$ לכל נקודה - גם בבעית האופט' וגם עבור החזאי שקיבלנו מפתרונה - דבר שתלוי בממד של $\mathbb{R}^{\tilde{d}}$, וזאת נוכל לבצע בעזרת **Kernel-Trick**: אשר מבצע את המכפלה: $K(\mathbf{x}', \mathbf{x}'') = \phi^T(\mathbf{x}') \cdot \phi(\mathbf{x}'')$ מבלי לעבור לממד החדש $\mathbb{R}^{\tilde{d}}$ אליו עוברים כאשר מחשבים את $\phi(\mathbf{x})$. כל **קרנל פולינומיאלי מהצורה**: $K(\mathbf{x}', \mathbf{x}'') = (\xi + c \cdot \mathbf{x}'^T\mathbf{x}'')^Q$ הוא קרנל ולידי. קרנל **ולידי נוסף** הוא **RBF**: $K(\mathbf{x}', \mathbf{x}'') = \exp\{-\gamma\|\mathbf{x}' - \mathbf{x}''\|^2\}$ שמחשב מכפלה פנימית במרחב אינסופי. **קרנל ולידי** הוא קרנל שניתן לרשום אותו בתור מכפלה פנימית מהצורה: $\phi^T(\mathbf{x}') \cdot \phi(\mathbf{x}'')$ עבור $\phi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$. כמו-כן K ולידי אם"ם : $\underline{1}$: סימטרי. $\underline{2}$: לכל סט נקודות $\mathbf{x}_1, \ldots, \mathbf{x}_N$ המטריצה שאבריה הם $K(\mathbf{x}_i, \mathbf{x}_j)$ היא P.S.D (mercer's).

התאמת-יתר: התאמת המודל למידע יותר מהנדרש - E_{in} טוב אבל E_{out} רע. הרבה פעמים מקבלים מדידות שמושתתות על רעש והתאמת יתר יכולה להתבטא בהתאמת המודל לאותם רעשים (נניח התאמת פולינום מסדר 4 עבור מדידות שמקורן בפולינום מסדר 2 עם קצת רעשים). הוספת דרגות חופש למודל שלנו יכולה להניב אמנם E_{in} מאוד נמוך (בזמן האימון) אך E_{out} מאוד גבוה (בזמן הטסט) כתוצאה מהרעשים - מה שיוביל להכללה גרועה של המודל. בצורה פשוטה: אם ישנם 2 חזאים: O ו- R (מלשון overfit ו- restricted) ושניהם מודעים לכך שפונקצית המטרה היא מסדר 10, O יבחר בסט ההיפותזות של פולינום מסדר 10: $\mathcal{H}_{10}$ ואילו R יבחר באלו מסדר 2: $\mathcal{H}_2$. השוני בבחירה יכול להתבטא בכך ש-O יותאם לרעשים במדידות ו-R יכליל את המגמה יותר טוב - משמע מה שחשוב **הוא לא פונקצית המטרה שלומדים** אלא: (1) כמות המדידות (2) איכות המדידות (רעש גדול או קטן). ההסבר לכך הוא שאם ישנן מס' קטן של נקודות מאוד רועשות שנוצרו ע"י פולינום ממעלה 10 - התאמה של פולינום אחר ממעלה 10 אליהן זוהי התאמת-יתר. **ככל שמס' המדידות יגדל כך ה** E_{in} **יתכנס לערך מסוים ואילו** E_{out} **ילך ויקטן.** את האפקט

של רעש וכמות המדידות נוכל לבחון בצורה פשוטה: נניח ופונקצית המטרה היא $y = f(x) + \varepsilon(x) = \sum_{q=0}^{Q_f} a_q x^q + \varepsilon(x)$ כאשר $\varepsilon(x)$ מייצג את רעש אקראי (נניח רעש גאוסי עם תוחלת 0 ושונות σ^2) ושישנן N נק'. אם נאמן שני מודלים, אחד $g_{10} \in \mathcal{H}_{10}$ ואחד $g_2 \in \mathcal{H}_2$ תמיד יתקיים ש- $E_{in}(g_{10}) \leq E_{in}(g_2)$ ואילו לא בהכרח ש- $E_{out}(g_{10}) < E_{out}(g_2)$ (ואף להיפך!), נבחן את ההפרש כדי לקבל את ה- overfit-mesaure: $E_{out}(g_{10}) - E_{out}(g_2)$. נסיק: $\underline{1}$: מס' נק' גדול מקטין את ה-overfitting. $\underline{2}$: רעש גדול מגדיל את ה-overfitting.

רגולריזציה: משמעותה הגבלת (צמצום) סט היפותזות קיים על מנת להתמודד עם overfitting. עבור סט תצפיות: $\mathbf{z} = [1, L_1(x), \ldots, L_2(x)]$ שבנוי מטרנס' על סט תצפיות אחר ונניח שקיים סט היפותזות $\mathcal{H}_Q = \left\{\sum_{q=0}^{Q} w_q L_q(x)\right\}$. אילוץ קשיח (או hard constraint) ידרוש איפוס של מקדמים מסוימים של ההיפותזות שנבדקות. משמע, אם מדובר בסט היפותזות של פול' ממעלה 10, הרי שפול' ממעלה 2 כלולים בתוך הסט הנ"ל ואם נרצה להגביל לפול' ממעלה 2 אז נגדיר אילוץ $w_q = 0$ לכל $q > 2$. מאידך קיימת האפשרות לאילוץ רך (או soft constraint) בו דורשים מקדמים קטנים, לדוגמה, שסכום ריבועי המקדמים יהיה קטן מערך קבוע: $\sum_{q=0}^{Q} w_q^2 \leq C$ ובכך מקטינים את מס' ההיפותזות האפשריות **(דרגות החופש)** ואת האפשרות להתאים את עצמנו לרעשים (כי בוחרים פונק' יותר חלקות). לפני רגולריזציה נהוג **לבצע סטנדרטיזציה לפיצ'רים** (החסרת הממוצע וחלוקה בסטיית התקן כדי לקבל שונות 1). ולאחר מכן להשתמש בשיטות שונות.

Ridge-Regression: מוגדרת בתור $\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} w_j x_{nj}\right)^2 \text{ s.t } \sum_{j=1}^{d} w_j^2 \leq C$ **או** לאחר שימוש בלגרנג'יאן על מנת להסיר את האילוצים: $\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} w_j x_{nj}\right)^2 + \lambda\sum_{j=1}^{d} w_j^2$ (עבור $\lambda \geq 0$) נסמן: $\bar{x}_j = \frac{1}{N}\sum_{n=1}^{d} x_{nj}$ (ממוצע ערכי פיצ'ר j) ונפתח את הביטוי האלגברי: $\sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} w_j\bar{x}_j - \sum_{j=1}^{d} w_j(x_{nj} - \bar{x}_j)\right)^2 + \lambda\sum_{j=1}^{d} w_j^2$ $(*)$ בנוסף, נסמן $w_0^c = w_0 - \sum_{j=1}^{d} w_j\bar{x}_j$. נגזור את הביטוי על פי w_0^c: $\frac{\partial}{\partial w_0^c} = -2\sum_{n=1}^{N}\left(y_n - w_0^c - \sum_{j=1}^{d} w_j(x_{nj} - \bar{x}_j)\right) = 0$ נזרוק קבועים ונקבל: $\sum_{n=1}^{N} y_n - Nw_0^c - \sum_{j=1}^{d} w_j\left(\sum_{n=1}^{N} x_{nj} - N\bar{x}_j\right) = 0$ ומהתאפסות הגורם האחרון נקבל: $\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n = w_0^c$. נציב חזרה ב-$(*)$ ונקבל: $\sum_{n=1}^{N}\left(y_n - \bar{y} - \sum_{j=1}^{d} w_j(x_{nj} - \bar{x}_j)\right)^2 + \lambda\sum_{j=1}^{d} w_j^2$ נסמן $\tilde{y}_n = y_n - \bar{y}$ ו- $\tilde{x}_{nj} = x_{nj} - \bar{x}_j$ ונקבל וקטור $\tilde{\mathbf{y}} = [\tilde{y}_1, \ldots, \tilde{y}_N]$ ואת המטריצה $\tilde{X} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_d]$ (העמודות הן וקטורי הפיצ'ר $1 \leq j \leq d$ ממורכזים) ונקבל: $\sum_{n=1}^{N}\left(\tilde{y}_n - \sum_{j=1}^{d} w_j\tilde{x}_{nj}\right)^2 + \lambda\sum_{j=1}^{d} w_j^2 = \left\|\tilde{\mathbf{y}} - \tilde{X}\mathbf{w}\right\|^2 + \lambda\|\mathbf{w}\|^2$ (שקול לבעית least squares עם פרמטר רגולריזציה). קיבלנו את הבעיה: $(1)\min_{\mathbf{w}\in\mathbb{R}^d} \left\|\tilde{\mathbf{y}} - \tilde{X}\mathbf{w}\right\|^2 + \lambda\|\mathbf{w}\|^2$. נפתח את הביטוי: $(\tilde{\mathbf{y}} - \tilde{X}\mathbf{w})^T(\tilde{\mathbf{y}} - \tilde{X}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w}$. כעת אפשר לעשות השלמה ריבועית כמו ברגרסיה לינארית או לגזור את הביטוי ולהשוות לאפס: $\frac{\partial}{\partial\mathbf{w}}(\cdot) = -2\tilde{X}^T\left(\tilde{\mathbf{y}} - \tilde{X}\mathbf{w}\right) + 2\lambda\mathbf{w} = \mathbf{0}$, נייפה, נזרוק קבועים ונקבל: $(**)(\tilde{X}^T\tilde{X} + \lambda I)\mathbf{w} = \tilde{X}^T\tilde{\mathbf{y}}$ (מזכיר את המשוואות הנורמליות למעט הביטוי של λI). בהנחה ש- $\lambda > 0$ - המטריצה $(\tilde{X}^T\tilde{X} + \lambda I)$ היא P.D: ניקח $\mathbf{c} \neq \mathbf{0}$ ונכפיל משמאל ומימין: $\mathbf{c}^T(\tilde{X}^T\tilde{X} + \lambda I)\mathbf{c} = \left\|\tilde{X}\mathbf{c}\right\|^2 + \lambda\|\mathbf{c}\|^2 \equiv (\geq 0 + > 0) > 0$ קיבלנו ביטוי חיובי (כי $\lambda > 0$ עבור רגולריזציה) ולכן הע"ע חיוביים **ממש** וקיבלנו **מטריצה הפיכה**! נכפיל את $(**)$ משמאל בהפכית ונקבל את $\mathbf{w}$: $(2)\mathbf{w} = (\tilde{X}^T\tilde{X} + \lambda I)^{-1}\tilde{X}^T\tilde{\mathbf{y}}$. ידוע כי $\hat{\tilde{\mathbf{y}}} = \tilde{X}(\tilde{X}^T\tilde{X} + \lambda I)^{-1}\tilde{X}^T\tilde{\mathbf{y}}$. נבחין כי $\lambda = 0$ מבטל את הרגולריזציה וכי λ גדול מידי יכול להוביל ל- underfitting ואף $\lambda \to \infty$ מאלץ $\mathbf{w} = \mathbf{0}$ כיוון שהמטריצה $(\tilde{X}^T\tilde{X} + \lambda I)$ תהיה אלכסונית עם אלכסון של ∞ ואילו ההפכית שלה תהיה $0 \equiv 1/\infty$ ואז נקבל $(2) = 0$ והחזאי שלנו יסתמך על $\bar{y}$ בלבד (ניתן להוכיח על ידי גזירת $\sum_{n=1}^{N}(y_n - w_0)^2$ והשוואה לאפס שמקבלים את $\bar{y}$). ה- λ האופטימלי מוגדר על פי הרעש (שונות) במדידות שלנו - רעש נמוך יצריך λ נמוך ורעש גבוה יצריך λ גבוה. (לדוגמה $\sigma^2 = 0$ מצריך $\lambda = 0$ ואילו $\sigma^2 = 0.5$ יכול להצריך $\lambda = 1$).

Lasso Regression: מוגדר בצורה דומה לרידג' למעט האילוץ: $\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}(y_n - w_0 - \sum_{j=1}^{d} w_j x_{nj})^2 \text{ s.t } \sum_{j=1}^{d} |w_j| \leq C$ ששקול ל- $\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} w_j x_{nj}\right)^2 + \lambda\sum_{j=1}^{d} |w_j|$ עבור $\lambda \geq 0$ (או $\lambda > 0$ עבור רגולריזציה). לבעיה זו אין פתרון סגור, אך אפשרי לנסח אותה בתור בעית QP. ההבדלים בין רידג' ללאסו: אמנם שניהם פותרים את בעית ה-least squares ומנסים לצמצם את מרחב ההיפותזות, אך לכל אחד אילוץ שונה, כאשר הלאסו בסבירות גבוהה יאפס חלק מהמקדמים (אברי הוקטור $\mathbf{w}$) לעומת הרידג' שלעתים קרובות פשוט מקטין את ערכי המקדמים בלבד. משמע לאסו מעודד פתרונות דלילים (מחיקת פיצ'רים) ובכך משמש גם ל-feature selection.

ולידציה: דרך נוספת להתמודדות עם overfitting והקטנת E_{out} (על פני הקטנת E_{in}) קבוצת אימון: תת קבוצה של התצפיות שעליה מאמנים את המודל ועל פיה מתקבל החזאי האופטימלי. קבוצת מבחן: תת-קבוצה של התצפיות **שלא נכללת** (!!) בקבוצת האימון ובתהליך האימון. משמע E_{test} הוא שערוך בלתי מוטה של E_{out}. קבוצת ולידציה: דומה לקבוצת מבחן בכך שלא משתמשים בה בזמן האימון של המודל אך משתמשים בה על מנת לבחור פרמטרים כאלו ואחרים שקשורים באימון המודל. ההשפעה של קבוצת הולידציה מספיק קטנה ולא מוטה, כדי שתספק לנו "תחושה" לגבי איכות המודל (E_{out}). ישנן שיטות שונות ליצירת קבוצות אימון, ולידציה ומבחן על פני N נקודות. K is taken out of N: **בהנחה שהדוגמאות בת"ס**, נגדיר $N-K$ דוגמאות לקבוצת האימון $\mathcal{D}_{train}$ שעל פיה נבחר חזאי אופטימלי g^-, ו- K דוגמאות לקבוצת ולידציה $\mathcal{D}_{val}$. (אסור שהחלוקה תהיה תלויה בערכי הדוגמאות). השגיאה תוגדר: $e(g^-(\mathbf{x}_n), y_n)$ ותניב: $\mathbb{E}_{\mathbf{x}_n}[e(g^-(\mathbf{x}_n), y_n)] = E_{out}(g^-)$ (תוחלת השגיאה על פני התפלגות $\mathbf{x}$ היא E_{out}) ושגיאת הולידציה: $E_{val}(g^-) = \frac{1}{K}\sum_{\mathbf{x}_n\in\mathcal{D}_{val}} e(g^-(\mathbf{x}_n), y_n)$. כמו-כן אם נרצה לחשב את התוחלת של $E_{val}(g^-)$ אז נחשב את: $\mathbb{E}_{\mathcal{D}_{val}}[E_{val}(g^-)] = \frac{1}{K}\sum_{\mathbf{x}_n\in\mathcal{D}_{val}} \mathbb{E}_{\mathcal{D}_{val}}[e(g^-(\mathbf{x}_n), y_n)] = \frac{1}{K}\sum_{\mathbf{x}_n\in\mathcal{D}_{val}} E_{out}(g^-) = E_{out}(g^-)$ (ולכן מדובר במשערך שאינו מוטה (ubiased) של $Eout$. את שונות השגיאה נגדיר: $\sigma^2(g^-) = Var_{\mathbf{x}}[e(g^-(\mathbf{x}_n), y_n)]$ ובצורה דומה נקבל: $\sigma_{val}^2 = Var_{\mathcal{D}_{val}}[E_{val}(g^-)] = \frac{1}{K}^2 Var(\sum_{\mathbf{x}_n\in\mathcal{D}_{val}} e(g^-(\mathbf{x}_n), y_n)) = \frac{1}{K}^2\sum_{\mathbf{x}_n\in\mathcal{D}_{val}} \sigma^2(g^-) = \frac{1}{K}\sigma^2(g^-)$ משמע עבור K גדול יותר נקבל שערוך מדויק יותר (כי נקטין את השונות). כמה נקודות: $\underline{1}$: עבור K גדול שונות השגיאה E_{val} קטנה עד נקודה בה K גדול מידי ואין מספיק דוגמאות ללמוד מהן (נוצר overfitting) $\underline{2}$: עבור K גדול השגיאה E_{val} של g^- גדלה (כי יש פחות דוגמאות ללמוד מהן). כלל אצבע: הקצאת $K = N/5$ מהדוגמאות לולידציה. רצוי ש-K יהיה מספיק גדול בשביל E_{val} אמין אך מספיק קטן כדי שיהיו מספיק דוגמאות מהן החזאי g^- יוכל ללמוד. $(*)$ בפועל מדווחים את $E_{val}(g^-)$ אך את המודל בפועל ניתן לאמן על כל N הדוגמאות, לקבל חזאי g ולהשתמש בו במקום ב- g^-. כמו-כן ניתן להשתמש בולידציה בשביל Model

Selection: הדרכה בעת תהליך הלמידה. אם נקבל מספר חזאים נוכל לבחון את טיבם בעזרת סט הולידציה (ולא על פי E_{in} שיכול לנבוע מ-overfitting) ולבצע בחירות מושכלות לגבי: 1: סוג המודל (לינארי? לא לינארי?) 2: דרגת הפולינום במודל 3: ערכי פרמטרי הרגולריזציה 4: כל בחירה אחרת שמשפיעה על תהליך הלמידה שלנו. בעצם, אם קיבלנו $\mathcal{H}_1 \dots \mathcal{H}_M$ היפותזות, נבחן אותן על פי סט הולידציה, נבחר את ההיפותזה עבורה E_{val} הוא הנמוך ביותר, נאמן אותה מחדש עם אותם פרמטרים על כל הדוגמאות שלנו וכך נשיג את המודל הסופי. Cross Validation: מצד אחד ככל ש-K קטן: $E_{out}(g) \approx E_{out}(g^-)$ ומצד שני ככל ש-K גדול: $E_{out}(g^-) \approx E_{val}(g^-)$ - הדילמה הנ"ל נפתרת על ידי Leave one out, על פי רעיון זה לכל $n = 1, \dots, N$ נבחר דוגמה $(\mathbf{x}_n, y_n)$ להיות לולידציה ואת שאר $N-1$ הדוגמאות לאימון, כך נקבל חזאים g_n^- שהשגיאה של אחד מהם: $e_n = E_{val}(g_n^-) = e(g_n^-(\mathbf{x}_n), y_n)$ והשגיאה הכללית היא מיצוע השגיאות הנ"ל: $E_{CV} = \frac{1}{N}\sum_{n=1}^N e_n$. על פי אותו רעיון ניתן להגדיר את k-fold CV שבו מוציאים קבוצה של דוגמאות במקום רק דוגמה אחת ומבצעים את אותה פרוצדורה.

רשתות נוירונים: שימוש במספר פרספטרונים מאפשר לפתור את בעית ה-XOR. כך קיבלו את ה- Multi-Layer Perceptron: מודל המורכב משכבות של נוירונים אשר כל אחד מהם יכול לקבל מספר קלטים, לחבר אותם ולהעבירם בפונקצית אקטיבציה שאינה בהכרח לינארית, כמו tanh שהיא חלקה וגזירה (לעומת $sign$) ודומה לפונקצית **הסיגמויד** ומוגדרת: $\tanh = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ כאשר: $\lim_{x\to\infty} = 1$ ו- $\lim_{x\to-\infty} = -1$ ונגזרתה: $1 - \tanh^2(x)$. הרשת מורכבת מ- L שכבות. השכבה הראשונה היא שכבת הקלט - Input Layer שהיא $\ell = 0$ **ולא נספרת מבין השכבות**. שכבת המוצא (פלט) - Output Layer שהיא $\ell = L$ וממנה יוצא החזאי, והשכבות החבויות - Hidden Layers שהן $0 < \ell < L$ ונמצאות בין שכבת הקלט ושכבת המוצא. **בכל שכבה חבויה** מוסיפים איבר נוסף - איבר הטיה ואין לו כניסות אך הוא מחובר לכל נוירון בשכבה הבאה למעט ה-bias הבא. הקלט של נוירון הוא סט הפלטים של השכבה הקודמת (בנוסף לאיבר ההטיה): $\mathbf{w}^T\mathbf{x} + b$. הסכום עובר דרך פונקצית האקטיבציה והמוצא המתקבל מהנוירון הוא $\theta(\mathbf{w}^T\mathbf{x} + b)$. עבור בעית רגרסיה נגדיר את פונקצית האקטיבציה של הנוירונים בשכבת הפלט כפונ' הזהות ונבדוק את השגיאה בעזרת שגיאה ריבועית. עבור בעית סיווג נגדיר את פונק' האקט' בתור $sign$ ואת השגיאה בתור שגיאה ריבועית או cross entropy. במקרה ורוצים הסתברות נשתמש בסיגמויד/softmax ובשגיאת cross entropy. לאחר שבחרנו את ארכיט' הרשת (מספר השכבות -L וכמה נוירונים בכל שכבה) ההיפותזה תוגדר על ידי בחירת משקלים עבור הרשת. סימונים: 1: $\mathbf{s}^{(\ell)}$ וקטור הכניסה לשכבה ℓ ממימד $d^{(\ell)}$ (אין $\mathbf{s}^{(0)}$) מוגדר בתור $W^{(\ell)}\mathbf{x}^{(\ell-1)}$. 2: $\mathbf{x}^{(\ell)}$ וקטור ממימד $d^{(\ell)} + 1$ שכולל את הפלט של נוירון בשכבה ℓ לאחר אקטיבציה. 3: $W^{(\ell)}$ מטריצת המשקולות שנכנסות לשכבה ℓ ממימד $(d^{(\ell-1)}+1) \times d^{(\ell)}$ שכל איבר $w_{i=\text{col},\, j=\text{row}}^{(\ell=\text{layer})}$ בה הוא המשקולת בין נוירון i בשכבה $\ell-1$ לנוירון j בשכבה ℓ (אין $W^{(0)}$). המעבר בין נוירון i עם יציאה $x_i^{(\ell-1)}$ (יוצא משכבה $\ell-1$) לנוירון j (בשכבה ℓ) מוגדר: $s_j^{(\ell)} = x_i^{(\ell-1)} \cdot w_{ij}^{(\ell)}$ וכלל ערכי $s_j^{(\ell)}$ הם הוקטור $(W^{(\ell)})^T\mathbf{x}^{(\ell-1)}$. התהליך השלם הוא: $\mathbf{x} = \mathbf{x}^{(0)} \xrightarrow{W^{(1)}} \mathbf{s}^{(1)} \xrightarrow{\theta} \mathbf{x}^{(1)} \xrightarrow{W^{(2)}} \mathbf{s}^{(2)} \xrightarrow{\theta} \mathbf{x}^{(2)} \dots \xrightarrow{W^{(L)}} \mathbf{s}^{(L)} \xrightarrow{\theta} \mathbf{x}^{(L)} = h(\mathbf{x})$. תהליך ה- forward propagation (*): 1: $\mathbf{x}^{(0)} \leftarrow \mathbf{x}$. 2: לכל $\ell = 1, \dots, L-1$ נגדיר: $\mathbf{s}^{(\ell)} \leftarrow (W^{(\ell)})^T\mathbf{x}^{(\ell-1)}$ ו- $\mathbf{x}^{(\ell)} \leftarrow \begin{bmatrix} 1 \\ \theta(\mathbf{s}^{(\ell)}) \end{bmatrix}$. 3: נחזיר את $\mathbf{x}^{(L)} = \theta((W^{(L)})^T\mathbf{x}^{(L-1)}) = h(\mathbf{x})$.

Backpropagation: נרצה לאפטם את $E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^N e_n$ (לרוב נשתמש בשגיאה ריבועית $\frac{1}{N}\sum_{n=1}^N (h(\mathbf{x}_n;\mathbf{w}) - y_n)^2$) על ידי שימוש ב-GD וחישוב של הנגזרת (מטריצה): $\frac{\partial E_{in}}{\partial W^{(\ell)}} = \frac{1}{N}\sum_{n=1}^N \frac{\partial e_n}{\partial W^{(\ell)}}$ לכל $\ell = 1, \dots, L$, אשר מורכבת מהרכיבים $\frac{\partial e}{\partial w_{ij}^{(\ell)}}$ בעזרת שימוש חוזר של **כלל השרשרת** לחישוב נגזרות חלקיות של שכבה ℓ ע"י הנגזרות החלקיות של שכבה $\ell + 1$: תחילה, כיוון שרק $\mathbf{s}_j^{(\ell)}$ **תלוי ב-** $w_{ij}^{(\ell)}$ נוכל לחשב: $\frac{\partial e}{\partial w_{ij}^{(\ell)}} = \frac{\partial \mathbf{s}_j^{(\ell)}}{\partial w_{ij}^{(\ell)}} \cdot \frac{\partial e}{\partial \mathbf{s}_j^{(\ell)}}$ אבל מכך ש- $\mathbf{s}_j^{(\ell)} = \sum_{\alpha=0}^{d^{(\ell-1)}} w_{\alpha j}^{(\ell)}\mathbf{x}_\alpha^{(\ell-1)}$ הרי ש- $\nabla_{w_{ij}}\mathbf{s}_j^{(\ell)} = \mathbf{x}_i^{(\ell-1)}$ ולכן: $\frac{\partial e}{\partial w_{ij}^{(\ell)}} = \mathbf{x}_i^{(\ell-1)} \cdot \frac{\partial e}{\partial \mathbf{s}_j^{(\ell)}}$. נסמן $\boldsymbol{\delta}_j^{(\ell)} = \frac{\partial e}{\partial \mathbf{s}_j^{(\ell)}}$ ובעצם עבור כל השכבה: $(1)\frac{\partial e}{\partial W^{(\ell)}} = \mathbf{x}^{(\ell-1)} \cdot (\boldsymbol{\delta}^{(\ell)})^T$. כעת נוכל לפתח את $\boldsymbol{\delta}^{(\ell)}$ שמגדיר את ה"רגישות" (sensitivity) באותה צורה: $\boldsymbol{\delta}_j^{(\ell)} = \frac{\partial e}{\partial \mathbf{s}_j^{(\ell)}}$ ורק $\mathbf{x}_j^{(\ell)}$ **מסתמך על** $\mathbf{s}_j^{(\ell)}$, נוכל להשתמש שוב בכלל השרשרת ולקבל: $\boldsymbol{\delta}_j^{(\ell)} = \frac{\partial e}{\partial \mathbf{s}_j^{(\ell)}} = \frac{\partial e}{\partial \mathbf{x}_j^{(\ell)}} \cdot \frac{\partial \mathbf{x}_j^{(\ell)}}{\partial \mathbf{s}_j^{(\ell)}}$ ידוע כי $\frac{\partial \mathbf{x}_j^{(\ell)}}{\partial \mathbf{s}_j^{(\ell)}} = \theta'(\mathbf{s}_j^{(\ell)})$ כי זו פונק' האקטיבציה ולכן: $\boldsymbol{\delta}_j^{(\ell)} = \theta'(\mathbf{s}_j^{(\ell)}) \cdot \frac{\partial e}{\partial \mathbf{x}_j^{(\ell)}}$. כעת נרצה לפתח מנגנון שיאפשר לחשב את $\boldsymbol{\delta}_j^{(\ell)}$ על פי $\boldsymbol{\delta}_j^{(\ell+1)}$ (שימוש בשכבה הבאה) ונבחין שהביטוי $\frac{\partial e}{\partial \mathbf{x}_j^{(\ell)}}$ יכול להיות מיוצג על ידי הנגזרות של e על פי $\mathbf{s}_k^{(\ell+1)}$ לכל נוירון k בשכבה הבאה (כי הרי כל נגזרת כזו מסתמכת על $\mathbf{x}_j^{(\ell)}$), משמע נוכל לרשום: $\frac{\partial e}{\partial \mathbf{x}_j^{(\ell)}} = \sum_{k=1}^{d^{(\ell+1)}} \frac{\partial \mathbf{s}_k^{(\ell+1)}}{\partial \mathbf{x}_j^{(\ell)}} \cdot \frac{\partial e}{\partial \mathbf{s}_k^{(\ell+1)}} = \sum_{k=1}^{d^{(\ell+1)}} w_{jk}^{(\ell+1)}\boldsymbol{\delta}_k^{(\ell+1)}$. ובסה"כ מצאנו: $\boldsymbol{\delta}_j^{(\ell)} = \theta'(\mathbf{s}_j^{(\ell)})\sum_{k=1}^{d^{(\ell+1)}} w_{jk}^{(\ell+1)}\boldsymbol{\delta}_k^{(\ell+1)}$ עבור $j = 1, \dots, d^{(\ell)}$.

וניתן לכתוב גם עבור כלל השכבה: $(2)\boldsymbol{\delta}^{(\ell)} = \theta'(\mathbf{s}_j^{(\ell)}) \otimes \left[W^{(\ell+1)}\boldsymbol{\delta}^{(\ell+1)}\right]_1^{d^{(\ell)}}$ ($\otimes$ מסמל הכפלה איבר-באיבר של מטריצות, בלי איבר 0 - ההטיה). את התהליך צריך להתחיל משכבה L ולכן צריך את $\boldsymbol{\delta}^{(L)}$ והוא כמובן מוגדר בתור הנגזרת של השגיאה (הריבועית במקרה שלנו) לפי $s^{(L)}$ (נזכור שזו השכבה האחרונה ולכן מדובר בסקלרים): $\boldsymbol{\delta}^{(L)} = \frac{\partial e}{\partial s^{(L)}} = \frac{\partial}{\partial s^{(L)}}(x^{(L)} - y)^2 = 2(x^{(L)} - y) \cdot \frac{\partial x^{(L)}}{\partial s^{(L)}} = 2(x^{(L)} - y)\theta'(s^{(L)})$ ולכן: $(3)\boldsymbol{\delta}^{(L)} = 2(x^{(L)} - y)\theta'(s^{(L)})$ ולדוגמה אם פונק' האקט' במוצא היא tanh אז: $\boldsymbol{\delta}^{(L)} = 2(x^{(L)} - y)(1 - (x^{(L)})^2)$. ואם פונק' הזהות נקבל: $\boldsymbol{\delta}^{(L)} = 2(x^{(L)} - y)$. תהליך ה-Backpropagation(**): 1: מריצים forward propagation על וקטור הקלטים $\mathbf{x}$ ועבור כל שכבה שומרים $\mathbf{x}^{(\ell)}$ ו- $\mathbf{s}^{(\ell)}$ על פי (*). 2: מאתחלים $\boldsymbol{\delta}^{(L)} \leftarrow (3)$. 3: לכל $\ell = L-1, \dots, 1$: נגדיר את הנגזרת המתאימה $\theta'(\mathbf{s}^{(\ell)})$ (אם $\theta = \tanh$ אז נגדיר: $(\theta'(\mathbf{s}^{(\ell)})) \leftarrow \left[1 - \mathbf{x}^{(\ell)} \otimes \mathbf{x}^{(\ell)}\right]_1^{d^{(\ell)}}$ ואז נחשב את $\boldsymbol{\delta}^{(\ell)}$ על פי (2). לאחר שיש לנו את $\boldsymbol{\delta}^{(\ell)}$ לכל ℓ, נוכל להשתמש ב- (1) על מנת לחשב את נגזרת השגיאה עבור כל אחד מהפרמטרים (מטריצות $W^{(\ell)}$) ובעצם לקבל את $\frac{\partial E_{in}}{\partial W^{(\ell)}} = \frac{1}{N}\sum_{n=1}^N \frac{\partial e_n}{\partial W^{(\ell)}}$ לכל ℓ ובעזרת שימוש ב- GD נעדכן את המשקלים של מטריצות $W^{(\ell)}$. תהליך האימון השלם של איטרציה יחידה (עבור שגיאה ריבועית ו-N תצפיות): 1: נאתחל $E_{in} = 0$ ולכל שכבה $\ell = 1, \dots, L$: מטריצת הגדריאנטים $G^{(\ell)} = 0 \cdot W^{(\ell)}$. 2: לכל תצפית $(\mathbf{x}_n, y_n)$ נבצע forward propagation לפי (*) ולאחריה backpropagation לפי (**). 3: נגדיר $E_{in} \leftarrow E_{in} + \frac{1}{N}(\mathbf{x}^{(L)} - y_n)^2$ 3.1: לכל $\ell = 1, \dots, L$ נגדיר את הגרדיאנט של E_{in} לפי $W^{(\ell)}$ על נקודה $\mathbf{x}_n$ בתור: $G^{(\ell)}(\mathbf{x}_n) = [\mathbf{x}^{(\ell-1)}(\boldsymbol{\delta}^{(\ell)})^T]$ ואז: $G^{(\ell)} \leftarrow G^{(\ell)} + \frac{1}{N}G^{(\ell)}(\mathbf{x}_n)$. בסיום האיטרציה נגדיר: $W^{(\ell)}(t+1) \leftarrow W^{(\ell)}(t) - \eta G^{(\ell)}(t)$ (GD)

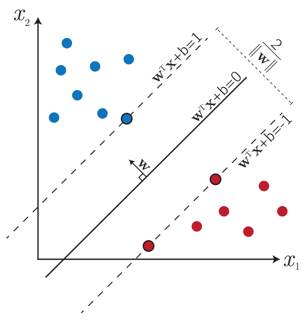
אבל נזכור כי **פונקצית המטרה מורכבת (לא קונבקסית) ולכן לא מובטחת התכנסות למינימום גלובלי ולכן** נאתחל את אלגוריתם GD ממספר נק' שונות (אתחולי $W^{(\ell)}$ שונים) כדי לנסות להמנע ממינימום לוקאלי ולבסוף להשוות בין החזאים השונים עם סט ולידציה. **נקודות חשובות**: ○ לא לאתחל $W^{(\ell)} = 0$ כי אז מהחישוב (2) נקבל $\boldsymbol{\delta}^{(\ell)} = 0$ ואז הנגזרות של פונקצית השגיאה הן 0 ולא נבצע שום שינוי. ○ אתחול של פרמטרים מאוד גדולים יכול לגרום לגדריאנט של פונקצית אקטיבציה כמו tanh להתאפס (vanishing gradient) ובכך $\theta' = 0$ ושוב מהחישוב (2) נקבל $\boldsymbol{\delta}^{(\ell)} = 0$ ולא נבצע עדכון. ○ ניתן להמנע מ- vanishing gradient עם שינוי פונקצית האקטיבציה במוצא לפונק' לינארית. ○ ניתן לאתחל את $W^{(\ell)}$ בצורה בת"ס עם מ"א בעל תוחלת 0 ושונות מאוד קטנה כך שהערכים הם סביב 0 אך לא ממש 0 (החלק הלינארי של tanh). ○ עוצרים את האלגוריתם כמו ב-GD רגיל. על מנת להתמודד עם overfitting נוכל להגדיל את מספר הנוירונים בשכבות החבויות ולהשתמש ברגולריזציה וולידציה. שיטה אחת לרגולריזציה נקראת Weight (squared) decay (שקול לרידג') ומוגדרת: $E_{aug}(\mathbf{w}) = E_{in}(\mathbf{w}) + \frac{\lambda}{N}\sum_{\ell,i,j}(w_{ij}^{(\ell)})^2$ ונגזרתה: $\nabla_{W^{(\ell)}}E_{aug}(\mathbf{w}) = \frac{\partial E_{in}(\mathbf{w})}{\partial W^{(\ell)}} + \frac{2\lambda}{N}W^{(\ell)}$ כאשר בשימוש ב-GD יגרום למשקלים להצטמצם עד כדי התאפסות. שיטה נוספת היא Weight elimination ומוגדרת: $E_{aug}(\mathbf{w}, \lambda) = E_{in}(\mathbf{w}) + \frac{\lambda}{N}\sum_{\ell,i,j}\frac{(w_{ij}^{(\ell)})^2}{1+(w_{ij}^{(\ell)})^2}$ ונגזרתה: $\nabla_{w_{ij}^{(\ell)}}E_{aug} = \frac{\partial E_{in}}{\partial w_{ij}^{(\ell)}} + \frac{2\lambda}{N} \cdot \frac{w_{ij}^{(\ell)}}{(1+(w_{ij}^{(\ell)})^2)^2}$ כאשר שימוש ב- GD יגרום למשקלים קטנים להיות עוד יותר קטנים ולבסוף מעלים אותם. שיטה נוספת לרגולריזציה היא Early stopping שמגבילה את מס' האיטרציות (ובהתאם גם את מרחב ההיפותזות) ומפסיקה את תהליך הלמידה לאחר t^* איטרציות ברגע שהשגיאה על סט הולידציה מתחילה לגדול. **ההבדל** העקרוני בין רשת נוירונים שמיוצגת על ידי $h(\mathbf{x}) = \theta(w_0 + \sum_{j=1}^m w_j\theta(\mathbf{v}_j^T\mathbf{x}))$ לבין פרספטרון שהקלט שלו עבר טרנס' לא לינארית: $h(\mathbf{x}) = \theta(w_0 + \sum_{j=1}^m w_j\phi(\mathbf{x}))$ הוא העובדה שפרספטרון קובע את $\phi(\mathbf{x})$ עוד לפני שהאלגוריתם לומד והוא פונ' לינ' של המשקלים ואילו רשתות נוירונים לומדות על פי $\mathbf{v}_j^T\mathbf{x}$ את הטרנס' הלא לינ' ומתאימות את אותם משקלים בזמן הלמידה כדי להניב פונ' לא לינ' של המשקלים.

למידה לא מונחית: Clustering: חלוקת קבוצת אובייקטים לתתי-קבוצות כך שכל האובייקטים באותה קבוצה (קלסטר) דומים/קרובים אחד לשני ואובייקטים בקלסטרים שונים - שונים אחד מהשני. הדמיון מוגדר על פי פונקצית דמיון/מרחק (דומה לפונקצית מטרה בלמידה מונחית) שהיא סימטרית ($d(x,y) = d(y,x)$) ומקיימת ש- $d(x,x) = 0$. אלגוריתמי קלסטרינג **מקבלים** סט אובייקטים ופונקצית דמיון והפלט הוא חלוקה של סט האובייקטים לתתי-קבוצות. K-means: חלוקה ל- K קלסטרים על פי סנטרואידים (מרכז כובד) כך שכל נקודה בקבוצה מסוימת קרובה לסנטראויד שלה יותר מלסנטרואידים אחרים. לאלגוריתם זה לא מובטחת ההתכנסות לאופט' גלובלי (מבחינה חישובית). האלגוריתם: בהינתן ערך $1 \le K \le N$, ונקודות: $\{\mathbf{x}_n\}_{n=1}^N$, כך ש- $\mathbf{x}_n \in \mathbb{R}^d$, נרצה למצוא וקטור סנטרואידים $\boldsymbol{\mu}_k \in \mathbb{R}^d$ $(k = 1, \dots, K)$ ווקטור שיוכים r_{nk} (שמקיים: אם $\mathbf{x}_n$ שייכת לקלסטר k אז: $r_{nk} = 1$ אחרת: $r_{nk} = 0$) שיביא למינימום את המרחקים בין כל נקודה לבין הסנטרואיד המתאים לה (המרחקים בתוך הקבוצות): $J = \sum_{n=1}^N\sum_{k=1}^K r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$. תחילה עבור $1 \le n' \le N$ נאפטם את $r_{n'k}$ (עבור n' כלשהו): $\min_{r_{n'k}} J = \min_{r_{n'k}}\sum_{k=1}^K r_{n'k}\|\mathbf{x}_{n'} - \boldsymbol{\mu}_k\|^2$, אנחנו נרצה להגדיר $r_{n'k} = 1$ עבור $\|\mathbf{x}_{n'} - \boldsymbol{\mu}_k\|^2$ מינימלי ולכן נוכל פשוט להגדיר: $j = \arg\min_k\|\mathbf{x}_{n'} - \boldsymbol{\mu}_k\|^2$ ואז להגדיר: $r_{n'j} = 1$ ולכל $k \neq j$: $r_{n'k} = 0$. כעת נאפטם לפי $\boldsymbol{\mu}_k$: $\min_{\boldsymbol{\mu}_\ell} J = \min_{\boldsymbol{\mu}_\ell}\sum_{n=1}^N r_{n\ell}\|\mathbf{x}_n - \boldsymbol{\mu}_\ell\|^2$, נגזור: $\frac{\partial}{\partial\boldsymbol{\mu}_\ell} = -2\sum_{n=1}^N r_{n\ell}(\mathbf{x}_n - \boldsymbol{\mu}_\ell) = 0$ ונקבל: $\sum_{n=1}^N r_{n\ell}\mathbf{x}_n = (\sum_{n=1}^N r_{n\ell})\boldsymbol{\mu}_\ell$, נחלץ את $\boldsymbol{\mu}_\ell$ ונקבל: $\boldsymbol{\mu}_\ell = \frac{\sum_{n=1}^N r_{n\ell}\mathbf{x}_n}{\sum_{n=1}^N r_{n\ell}}$ לכל $\ell = 1, \dots, K$. המכנה של התוצאה הוא בעצם סכימת מספר הנקודות בקלסטר ℓ ונוכל לסמנו N_ℓ. המונה משמעו סכימת ערכי $\mathbf{x}_n$ במידה והוא שייך לקלסטר ℓ משמע קיבלנו חלוקה של **סכום** הנקודות בקלסטר ℓ **במספר** הנקודות בקלסטר ℓ וזה בעצם ממוצע של הנקודות ששייכות לקלסטר ℓ - מצאנו $\boldsymbol{\mu}_\ell$ אופטימלי והוא לא בהכרח אחת מהדוגמאות. בשלב הבא נשייך לכל נקודה קלסטר על פי המרחק מהממוצעים החדשים שמצאנו ונחזור על התהליך שוב ושוב עד אשר לא יהיו יותר שינויים (חישבנו ממוצעים וההשמה של הנקודות לסנטרואידים לא שינתה את הקלסטרים) או עד מס' איטרציות מוגדר. אתחול: על פי בחירת נציגים מתוך הדוגמאות בצורה אקראית כסנטרואידים (kmeans++ בוחר את הסנטרואיד הראשון באקראי ואז כל סנטרואיד נוסף בוחר באופן פרופרציונלי למרחק מהסנטרואידים הקודמים (ככל שרחוק הסתברות גבוהה יותר). במידה ובוחרים פונקצית מרחק שונה (לא ריבועית, שכן רגישה לאאוטליירס) בד"כ מאלצים את הסטנרואידים להיות נקודות מתוך הדוגמאות. **נהוג** לאתחל בצורות שונות ולבחור את הטוב מבין כולם. ראוי לבחור K שהגדלה שלו לא מפחיתה באופן משמעותי את J. PCA: מציאת סט פיצ'רים ממימד נמוך יותר אשר מייצג נאמנה את הבעיה שלנו על ידי מחיקת ממדים יתירים. באופן שקול לכך ששימוש בטרנס' לא לינ' מגדיל ממד, מקטין E_{in} **ופוגע בהכללה**, הורדת ממד עלולה להגדיל E_{in} אך **תשפר את ההכללה**. 2 הגדרות: 1: הטלה אורתוג' של המידע לתת-מרחב מממד נמוך כך שהשונות של המידע המוטל היא מקס' 2: ההטלה הלינארית שמביאה למינ' את ממוצע ריבועי המרחקים בין התצפיות המקוריות לבין המוטלות. הגדרה 1: בהינתן $\{\mathbf{x}_n\}_{n=1}^N$ כך ש-$\mathbf{x}_n \in \mathbb{R}^D$ נרצה להטיל על ממד $M < D$ תוך מקסום השונות. עבור $M = 1$: נבחר $\mathbf{u}_1 \in \mathbb{R}^D$ כאשר $\mathbf{u}_1^T\mathbf{u}_1 = 1$(*). ההטלה של $\mathbf{x}_n$ על $\mathbf{u}_1$ היא: $\mathbf{z}_n = \frac{\mathbf{x}_n^T\mathbf{u}_1}{\mathbf{u}_1^T\mathbf{u}_1}\mathbf{u}_1$ נחשב את הממוצע על מנת לחשב את השונות של הנקודות המוטלות (ללא כיוון): ממוצע: $\frac{1}{N}\sum_{n=1}^N\mathbf{x}_n^T\mathbf{u}_1 = (\frac{1}{N}\sum_{n=1}^N\mathbf{x}_n)^T\mathbf{u}_1 = \overline{x}^T\mathbf{u}_1$, והשונות: $\frac{1}{N}\sum_{n=1}^N(\mathbf{x}_n^T\mathbf{u}_1 - \overline{x}^T\mathbf{u}_1)^2 = \frac{1}{N}\sum_{n=1}^N((\mathbf{x}_n - \overline{x})^T\mathbf{u}_1)^2$ מכך ש- $(\mathbf{x}_n - \overline{x})^T\mathbf{u}_1$ סקלר נוכל להפוך את המכפלה ונקבל: $\frac{1}{N}\sum_{n=1}^N((\mathbf{x}_n-\overline{x})^T\mathbf{u}_1)^2 = \frac{1}{N}\sum_{n=1}^N\mathbf{u}_1^T(\mathbf{x}_n-\overline{x})(\mathbf{x}_n-\overline{x})^T\mathbf{u}_1 = \mathbf{u}_1^T[\frac{1}{N}\sum_{n=1}^N(\mathbf{x}_n-\overline{x})(\mathbf{x}_n-\overline{x})^T]\mathbf{u}_1$ והביטוי בתוך הסוגריים המרובעים הוא מטריצת ה-covariance (שהיא סימטרית) של המדידות $\{\mathbf{x}_n\}_{n=1}^N$, נסמנה ב-S ואת כלל הביטוי ב- $J = \mathbf{u}_1^TS\mathbf{u}_1$. כיוון שנתון (*), נחפש את כיוון $\mathbf{u}_1$ בו השונות תהיה מקס': $(1)\ \max_{\mathbf{u}_1}\mathbf{u}_1S\mathbf{u}_1 \text{ s.t } \|\mathbf{u}_1\|^2 = 1$ נשתמש בכופל לגרנג' (לאחר תיקון האילוץ): $\tilde{J} := \mathbf{u}_1^TS\mathbf{u}_1 + \lambda(1 - \mathbf{u}_1^T\mathbf{u}_1)$, $\max_{\mathbf{u}_1}$, נגזור לפי $\mathbf{u}_1$: $\frac{\partial\tilde{J}}{\partial\mathbf{u}_1} = (S + S^T)\mathbf{u}_1 - 2\lambda\mathbf{u}_1 = 0 \overset{S^T=S}{\iff} S\mathbf{u}_1 = \lambda\mathbf{u}_1$ ולכן $\mathbf{u}_1$ הוא ו"ע של S ואם נכפיל משמאל ב-$\mathbf{u}_1$ נקבל: $\mathbf{u}_1^TS\mathbf{u}_1 = \lambda$ משמע כדי למקסם את (1) נגדיר את $\mathbf{u}_1$ להיות הו"ע עם ע"ע הגדול ביותר של מטריצה S. באופן כללי קיבלנו: $S\mathbf{u}_i = \lambda_i\mathbf{u}_i, i = 1, \dots, D$ ובצורה מטריציונית, אם נגדיר $U =$(מטריצה שעמודותיה הם $\mathbf{u}_i$) ו- $\Lambda = diag(\lambda_1, \dots, \lambda_D)$ אז: $SU = U \cdot \Lambda$ ומהפירוק הספקטרלי: $S = U\Lambda U^T$. משמע- ההטלה למרחב מממד M שתמקסם את השונות מוגדרת ע"י M הו"ע המתאימים לע"ע הגדולים ביותר של מטריצה S. ועבור הגדרה 2: נגדיר בסיס אורתונורמלי חדש $B = (\mathbf{u}_1, \dots, \mathbf{u}_D)$ ל $\mathbb{R}^D$ ונוכל לרשום לכל n: $\mathbf{x}_n = \sum_{i=1}^D(\mathbf{x}_n^T\mathbf{u}_i)\mathbf{u}_i$ ואז נקרב את הנקודה ב- M ממדים: $\tilde{\mathbf{x}}_n = \sum_{i=1}^M z_{ni}\mathbf{u}_i + \sum_{i=M+1}^D b_i\mathbf{u}_i$ (לא צריך את האיבר השני (bias) אם המידע ממורכז) ונרצה להביא למינימום את השגיאה בין $\mathbf{x}_n$ ל- $\tilde{\mathbf{x}}_n$: $J = \frac{1}{N}\sum_{n=1}^N\|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2 = \frac{1}{N}\sum_{n=1}^N\left\|\mathbf{x}_n - (\sum_{i=1}^M z_{ni}\mathbf{u}_i + \sum_{i=M+1}^D b_i\mathbf{u}_i)\right\|^2$ נגזור לפי z_{mj} ו- b_j: $\nabla_{z_{mj}} = \frac{2}{N}(\mathbf{u}_j^T\mathbf{x}_m - z_{mj}) = 0$ ו- $\nabla_{b_j} = \frac{2}{N}\sum_{n=1}^N(\mathbf{u}_j^T\mathbf{x}_n - b_j) = 0$ ונקבל את הפתרונות: $z_{nj} = \mathbf{u}_j^T\mathbf{x}_n$ עבור

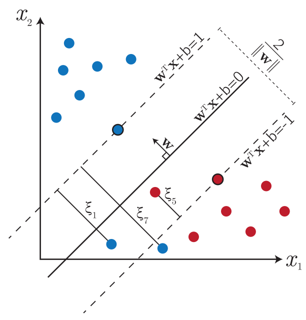
$j = 1,\dots,M$ ו- $b_j = \mathbf{u}_j^T\bar{\mathbf{x}} = \mathbf{u}_j^T\frac{1}{N}\sum_{n=1}^N\mathbf{x}_n$ עבור $j = M+1,\dots,D$ השגיאה לאחר ההצבה תניב לנו – $\mathbf{x}_n - \tilde{\mathbf{x}}_n = \sum_{i=1}^D(\mathbf{x}_n^T\mathbf{u}_i)\mathbf{u}_i - \sum_{i=1}^M(\mathbf{x}_n^T\mathbf{u}_i)\mathbf{u}_i - \sum_{i=M+1}^D(\bar{\mathbf{x}}^T\mathbf{u}_i)\mathbf{u}_i = \sum_{i=M+1}^D[(\mathbf{x}_n-\bar{\mathbf{x}})^T\mathbf{u}_i]\mathbf{u}_i$ נבחין שהשגיאה שקיבלנו אורתוג' למרחב עליו הטלנו את המדידות (כי פיצלנו מרחב אורתונ') ואז נגדיר $J = \frac{1}{N}\sum_{n=1}^N(\sum_{i=M+1}^D[(\mathbf{x}_n-\bar{\mathbf{x}})^T\mathbf{u}_i]\mathbf{u}_i)^T(\sum_{i=M+1}^D[(\mathbf{x}_n-\bar{\mathbf{x}})^T\mathbf{u}_i]\mathbf{u}_i)$ מאורתונורמליות $\mathbf{u}_i$ נקבל: $J = \frac{1}{N}\sum_{n=1}^N\sum_{i=M+1}^D[(\mathbf{x}_n-\bar{\mathbf{x}})^T\mathbf{u}_i]^2 = \sum_{i=M+1}^D\mathbf{u}_i^TS\mathbf{u}_i$ ונקבל בעית אופט' הפוכה להגדרה 1: נרצה למצוא $D-M$ ו"ע עם ע"ע נמוכים ביותר ונטיל את המידע על הוקטורים הנותרים. סה"כ: J האופטימלי מוגדר: עבור הגדרה 1: $\sum_{i=1}^M\lambda_i$ כאשר λ_i ע"ע הכי גדולים. הגדרה 2: $\sum_{i=M+1}^D\lambda_i$ כאשר λ_i ע"ע הכי קטנים. Data pre-processing: בהינתן מטריצה X עם N דוגמאות (שורות) ו- d פיצ'רים (עמודות) נוכל לבצע טרנס' מוכרות בשביל התהליך אשר יניבו $\mathbf{z}_n = \phi(\mathbf{x}_n)$ טוב יותר: 1: מרכוז - centering: $\mathbf{z}_n = \mathbf{x}_n - \bar{\mathbf{x}}$ כאשר $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^N\mathbf{x}_n$ 2: נרמול: $\frac{\mathbf{x}_n-\bar{\mathbf{x}}}{\sigma}$ (σ – סטיית תקן) בלי שינוי קורלציות. 3: whitening: נרמול עם דה-קורלציה על ידי $\mathbf{z}_n = \Lambda^{-1/2}U^T(\mathbf{x}_n-\bar{\mathbf{x}})$ (Λ ו- U מהפירוק של הגדרה 1). את החישובים הללו מבצעים על דוגמאות האימון **בלבד** ומשתמשים בהם גם באימון וגם בטסט.

שערוך: בהינתן סט מדידות $\mathbf{y}$ רוצים לשערך את ערכי $\mathbf{x}$ אשר קשורים לערכי $\mathbf{y}$. שתי גישות: 1: הגישה הפרמטרית - $\mathbf{x}$ הוא סט לא אקראי של פרמטרים דטרמנסטיים לא ידועים 2: הגישה הבייסיאנית - $\mathbf{x}$ הוא וקטור אקראי. גישה פרמטרית: x הוא פרמטר לא אקראי. נגדיר את השערוך של x על פי y בתור: $\hat{x}(y)$, משמע השגיאה: $e(y) = \hat{x}(y) - x = \hat{x} - x$. **חשוב לזכור** שאסור ש $\hat{x}$ יהיה מבוסס על ערכי x. מכך שאנחנו לא יודעים את ערכי $\mathbf{x}$ אבל יודעים את ערכי $\mathbf{y}$ (שהם מדידות רועשות), נגדיר את $\mathbf{y}$ בתור וקטור אקראי ונוכל להשתמש בו לחישוב ההטיה כאשר נרצה שהמשערך יהיה **חסר הטיה, כלומר שתוחלת השגיאה של המשערך תהיה** 0. את ההטיה נגדיר: $b_{\hat{x}}(x) = \mathbb{E}(e(y)) = \mathbb{E}(\hat{x}(y)) - x$ ונסיק שמשערך חסר הטיה מניב: $\mathbb{E}(\hat{x}(y)) = x$. כמו-כן נרצה לחשב את מטריצת ה-cov של השגיאות, כלומר נרצה להגדיר את $\Lambda_e(x) = \mathbb{E}\left[(e(y)-\mathbb{E}(e(y)))\cdot(e(y)-\mathbb{E}(e(y)))^T\right]$ (המטריצה המתקבלת ממכפלת הפרשי השגיאות מהתוחלת שלהן). maximum likelihood: משערך ששואף לערך האמיתי של הפרמטרים ככל שישנן יותר מדידות (תכונה אסימפטוטית), כלומר, ככל שיהיו יותר מדידות ההטיה תשאף לאפס והשונות סביב הפרמטר האמיתי תשאף לאפס ונוכל לקרוא למשערך **עקבי**. כפי שהזכרנו, $\mathbf{y}$ הוא וקטור אקראי שמתבסס (בחלקו) על $\mathbf{x}$ ולכן יש לו פונ' צפיפות: $f_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ (; עבור תלות בפרמטר דטרמיניסטי). את המשערך נסמן ב- $(1)L(\mathbf{x}) = f_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ כאשר $\mathbf{y}$ קבוע ונרצה למצוא $\mathbf{x}$ שימקסם את L: $\hat{\mathbf{x}}_{ML} = \arg\max_{\mathbf{x}}L(\mathbf{x})$. במקרים רבים קל יותר למקסם פונ' מונוט' ולכן **נרכיב** log על L ונמקסם את: $\hat{\mathbf{x}}_{ML} = \arg\max_{\mathbf{x}}\log L(\mathbf{x})$ ומתכונת המונוט' אותו ה-$\mathbf{x}$ ימקסם גם את L. במקרים בהם L גזירה ניתן למצוא מקס' על פי גזירה (בין אם של L או $\log L$) ובמידת הצורך (אם אין פתרון סגור או שקשה למצוא פתרון) להשתמש ב-GD ונגזרותיו. דוגמה 1: הטלת מטבע עם הסתברות עץ θ ופאלי $1-\theta$, ישנן N מדידות iid מתפלגות ברנולי. נרצה לשערך את θ. נוכל לרשום: $f_{y_n}(y_n;\theta) = \theta^{y_n}\cdot(1-\theta)^{1-y_n}$ עבור $y_n\in\{0,1\}$, ולכן עבור כל התצפיות $\mathbf{y}$ (שהן בת"ס): $f_{\mathbf{y}}(\mathbf{y};\theta) = \prod_{n=1}^N f_{y_n}(y_n;\theta) = \prod_{n=1}^N\theta^{y_n}\cdot(1-\theta)^{1-y_n} = \theta^{\sum_{n=1}^N y_n}\cdot(1-\theta)^{N-\sum_{n=1}^N y_n}$. נרצה למצוא מקס', כדי לפשט נפעיל את log: $\log f_{\mathbf{y}}(\mathbf{y};\theta) = \sum_{n=1}^N y_n\cdot\log\theta + (N-\sum_{n=1}^N y_n)\cdot\log(1-\theta)$ נמצא מקס' עם גזירה: $\nabla_\theta\log f_{\mathbf{y}}(\mathbf{y};\theta) = \frac{1}{\theta}(\sum_{n=1}^N y_n) + (N-\sum_{n=1}^N y_n)\cdot\frac{-1}{(1-\theta)} = 0$ ונקבל: $\frac{1}{\theta(1-\theta)}\cdot[(\sum_{n=1}^N y_n)(1-\theta) - (N-\sum_{n=1}^N y_n)\cdot\theta] = 0 \iff \frac{1}{\theta(1-\theta)}\cdot[(\sum_{n=1}^N y_n) - N\theta] = 0 \iff \frac{N}{\theta(1-\theta)}\cdot[\frac{1}{N}\sum_{n=1}^N y_n - \theta] = 0$ נחלץ את θ ונקבל את הערך האופטימלי: $\hat{\theta}_{ML} = \bar{y} = \frac{1}{N}\sum_{n=1}^N y_n$ (מתכנסים לממוצע האמפירי על פי חוק המספרים הגדולים). ואכן התוחלת של המשערך האופטימלי: $\mathbb{E}(\hat{\theta}_{ML}) = \frac{1}{N}\sum_{n=1}^N\mathbb{E}(y_n) \overset{\text{bernoulli}}{=} \theta$ ותוחלת השגיאה: $\mathbb{E}(e(\mathbf{y})) = \mathbb{E}(\hat{\theta}_{ML}(\mathbf{y})) - \theta = \theta - \theta = 0$ ולכן זהו משערך חסר הטיה. דוגמה 2: עבור התפלגות גאוסית, $f_{y_n}(y_n;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\{-\frac{1}{2\sigma^2}(y_n-\mu)^2\}$ וערכי y_n בת"ס, כמו-כן: $f_{\mathbf{y}}(\mathbf{y};\mu,\sigma^2) = \prod_{n=1}^N f_{y_n}(y_n;\mu,\sigma^2) = (2\pi\sigma^2)^{-N/2}\exp\{-\frac{1}{2\sigma^2}\sum_{n=1}^N(y_n-\mu)^2\}$ נפעיל log: $\log f_{\mathbf{y}}(\mathbf{y};\mu,\sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^N(y_n-\mu)^2$. נוכל לגזור לפי μ ולפי σ^2 ולקבל: $\nabla_\mu \Rightarrow \hat{\mu}_{ML} = \bar{y}$ ו- $\nabla_{\sigma^2} \Rightarrow \widehat{\sigma^2}_{ML} = \frac{1}{N}\sum_{n=1}^N(y_n-\bar{y})^2$. **הערה חשובה**: יש צורך לגזור פעם שנייה ולוודא שאכן הנקודה היא מקס', אם יש יותר ממשתנה אחד (כמו במקרה הגאוסי) יש צורך לוודא שההסיאן הוא N.D. גישת ML וריבועים פחותים: נניח: $\mathbf{y} = X\mathbf{w}+\mathbf{v}$ עבור $\mathbf{w}\in\mathbb{R}^{d+1}$ ו- $\mathbf{v} = [v_1,\dots,v_N]$ כאשר $v_i \sim N(0,\sigma^2)$ ו- v_i בת"ס, משמע $\mathbf{v}\sim N(\mathbf{0},\sigma^2\cdot I)$ ולכן כיוון ש-X קבוע, התוחלת: $\mathbb{E}(X\mathbf{w}+\mathbf{v}) = X\mathbf{w}+\mathbb{E}(\mathbf{v}) = X\mathbf{w}$, והשונות לא משתנה (הזזה בקבוע) ולכן: $\mathbf{y}\sim N(X\mathbf{w},\sigma^2I)$ ולכן: $f_{\mathbf{y}}(\mathbf{y};X\mathbf{w},\sigma^2I) = (\sqrt{\det(2\pi\sigma^2I)})^{-1}\exp\{-\frac{1}{2\sigma^2}\|\mathbf{y}-X\mathbf{w}\|^2\}$ נעבור log תוך הוצאת הקבוע מהdet: $\log f = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y}-X\mathbf{w}\|^2$ ובעצם $\hat{\mathbf{w}}_{ML} = \arg\max_{\mathbf{w}}\log(f_{\mathbf{y}}) = \arg\min_{\mathbf{w}}\|\mathbf{y}-X\mathbf{w}\|^2 = \hat{\mathbf{w}}_{LS}$ (זהה לשיטת הריבועים הפחותים). גישת ML ורגרסיה לוגיסטית: מתכונת הסיגמויד נוכל להגדיר: $p(y_n\mid\mathbf{x}_n) = \theta(y_n\mathbf{w}^T\mathbf{x}_n)$ ולכן: $L = \prod_{n=1}^N p(y_n\mid\mathbf{x}_n) = \prod_{n=1}^N\theta(y_n\mathbf{w}^T\mathbf{x}_n)$ נפעיל log: $\log(L) = \sum_{n=1}^N\log\theta(y_n\mathbf{w}^T\mathbf{x}_n)$ מכך ש $\log(1/x) = -\log x$ נקבל: $\log(L) = -\sum_{n=1}^N\log(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n})$ ונרצה: $\hat{\mathbf{w}}_{ML} = \arg\max_{\mathbf{w}}\log L(\mathbf{w}) = \arg\min_{\mathbf{w}} -L(\mathbf{w})$ והכפלה בקבוע לא תשנה את התוצאה ולכן: $\hat{\mathbf{w}}_{ML} = \arg\min_{\mathbf{w}}\frac{1}{N}\sum_{n=1}^N\log(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n})$ וזו בדיוק פונקצית המחיר שרצינו לאפטם ברגרסיה לוגיסטית (קרוס אנטרופי). גישה בייסיאנית: מתייחסת ל $\mathbf{x}$ כוקטור אקראי **וניתנת לנו פונקצית פילוג א-פריורי (לפני הגעת תצפיות) שלו** שהיא $p_{\mathbf{x}}(\mathbf{x})$, בנוסף גם נוכל להסיק את $p_{\mathbf{y},\mathbf{x}}(\mathbf{y},\mathbf{x}) = p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}\mid\mathbf{x})p_{\mathbf{x}}(\mathbf{x})$ (התפלגות משותפת של $\mathbf{y}$ ו- $\mathbf{x}$) ובנוסף **על פי חוק בייס** נסיק את הפילוג **הא-פוסטריורי** (איך מושפע פילוג $\mathbf{x}$ לאחר הגעת תצפיות): $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})$. על מנת למצוא את המשערך הטוב ביותר נגדיר פונקצית מחיר בין $\mathbf{x}$ למשערך שלנו $\mathbf{f}(\mathbf{y})$ כך: $C(\mathbf{x},\mathbf{f}(\mathbf{y}))$ וננסה להביא למינ' את התוחלת של פונקצית המחיר על פי בחירת המשערך $\mathbf{f}$. את המשערך האופטימלי נסמן ב- $\hat{\mathbf{x}}$ ובעית האופט' המלאה: $(2)\hat{\mathbf{x}}(\cdot) = \arg\min_{\mathbf{f}(\cdot)}E(C(\mathbf{x},\mathbf{f}(\mathbf{y})))$. מכך שגם $\mathbf{x}$ וגם $\mathbf{y}$ הם משתנים אקראיים נקבל ש: $\mathbb{E}(C(\mathbf{x},\mathbf{f}(\mathbf{y}))) = \int\int C(\mathbf{x},\mathbf{f}(\mathbf{y}))\cdot p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} = \int[\int c(\mathbf{x},\mathbf{f}(\mathbf{y}))p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x}]p_{\mathbf{y}}(\mathbf{y})d_{\mathbf{y}}$ וכדי להביא למינ' את הביטוי הנ"ל, נוכל לסמן $\mathbf{a} = \mathbf{f}(\mathbf{y})$ ולהביא למינ' את הביטוי הפנימי: $(3)\hat{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{a}}\int C(\mathbf{x},\mathbf{a})p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x}$ נשתמש בבייס ונשמיט את המכנה (כי הוא קבוע ביחס לאינטגרל) נקבל: $\hat{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{a}}\int C(\mathbf{x},\mathbf{a})p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}\mid\mathbf{x})p_{\mathbf{x}}(\mathbf{x})d\mathbf{x}$. Minimum Uniform Cost (MUC): נגדיר פונ' מחיר $C(a,\hat{a}) = 1 \text{ if } |a-\hat{a}|>\varepsilon, \text{ else } 0$. נציב את C ב-(3) ומכך שהפונקציה היא 1 רק עבור $|a-\hat{a}|>\varepsilon$ נסתכל על $\int$ רק עבור תנאי זה ונקבל: $\hat{x}_{MUC}(\mathbf{y}) = \arg\min_a[\int_{\{x|c=1\}}p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x}]$ כמו-כן בצורה שקולה אפשרי לחשב המאורע בו: $|a-\hat{a}|\le\varepsilon$ ולקבל: $\arg\min_a[1-\int_{a-\varepsilon}^{a+\varepsilon}p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x}]$ ובצורה שקולה: $\hat{x}_{MUC}(\mathbf{y}) = \arg\max_a[\int_{a-\varepsilon}^{a+\varepsilon}p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x}$ - כלומר, עבור כל נקודה a על פונקצית הפילוג (א-פוסטריורי) המשערך מגדיר סביבת ε של a ומבצע אינטגרציה בתחום שהתקבל סביב a מה שמתפרש בתור **ההסתברות של** $\hat{a}$ **להיות בסביבת** ε **של** a **בהינתן** $\mathbf{y}$. **ואם נשאיף** את ε לאפס המשערך יחפש את **הנק' המקס'** של פונק' הפילוג הא-פוסטריורית: Maximum A-Posterirori

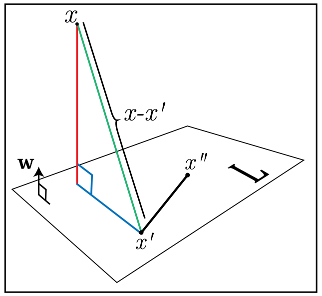
(MAP) Estimation, בעיה זו מוגדרת: $\hat{x}_{MAP}(\mathbf{y}) = \arg\max_a p_{x|\mathbf{y}}(a\mid\mathbf{y}) = \lim_{\varepsilon\to0}\hat{x}_{MUC}(\mathbf{y})$. פיתוח זה לאחר שימוש בחוק בייס, מוגדר: $(4)\hat{x}_{MAP}(y) = \arg\max_a\log p_{y|x}(y\mid a) + \log p_x(a)$ ושונה מגישת ML שמוגדרת: $\hat{x}_{ML} = \arg\max_a p(y;a) = \arg\max_a\log p(y;a)$ בכך שנוסף איבר $\log p_x(a)$ שנותן מידע א-פריורי לגבי הפרמטר אותו רוצים לשערך. Minimum Absolute Error (MAE): בהגדרת פונ' מחיר $C(a,\hat{a}) = |a-\hat{a}|$ נקבל $\hat{\mathbf{x}}(\mathbf{y}) = \arg\min_a[\int_{-\infty}^{\infty}|x-a|\,p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x}]$ והערך (**יכול להיות יותר מאחד**) שיקיים את השוויון יהיה **החציון:** $\int_{-\infty}^{\hat{x}_{MAE}(\mathbf{y})}p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x} = \int_{\hat{x}_{MAE}(\mathbf{y})}^{\infty}p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})d\mathbf{x} = 0.5$. Minimum Mean Square Error (MMSE): משערך **התוחלת המותנית** (ממוצע של x אחרי שראינו את $\mathbf{y}$), יחיד וקונבקסי, פונק' המחיר: $C(\mathbf{a},\hat{\mathbf{a}}) = \|\mathbf{a}-\hat{\mathbf{a}}\|^2$ ומתקבל ממנו $\hat{\mathbf{x}}_{MMSE}(\mathbf{y}) = \int_{-\infty}^{\infty}\mathbf{x}p(\mathbf{x}\mid\mathbf{y})dx = \mathbb{E}(\mathbf{x}\mid\mathbf{y})$. משערך זה חסר הטיה: $b_{MMSE} = \mathbb{E}[\mathbf{e}(\mathbf{x},\mathbf{y})] = \mathbb{E}[\hat{\mathbf{x}}_{MMSE}(\mathbf{y})-\mathbf{x}] = \mathbb{E}_{\mathbf{y}}[\mathbb{E}_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})]-\mathbb{E}[\mathbf{x}] = \mathbf{0}$, מטריצת ה-cov: $\Lambda_{MMSE} = \mathbb{E}[\mathbf{e}\mathbf{e}^T] = \mathbb{E}[\Lambda_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})] = \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}\mathbf{y}]) - (\mathbf{x}-\mathbb{E}[\mathbf{x}\mid\mathbf{y}])^T\right]$ ושגיאת השערוך אורתוגונלית לכל פונקציה של המדידות: $\mathbb{E}[(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})\mathbf{g}^T(\mathbf{y})] = \mathbf{0}$. **כלל המשערכים** צריכים את פונקצית הפילוג הא-פוסטריורית (התפלגות x לאחר צפייה ב- $\mathbf{y}$). **הערה**: במקרה ש $\mathbf{y}$, $\mathbf{x}$ מפולגים גאוסית במשותף ($p(\mathbf{x}\mid\mathbf{y})$ גאוסי) - התוצאות של MAE, MMSE, MAP מתכנסות לאותו ערך. משערך שמתלכד עם Ridge Regression: בהינתן סט מדידות $y_n = w_0 + \mathbf{x}_n^T\mathbf{w} + V_n$ כאשר: $\mathbf{x}_n$ תצפיות, $V_n\sim N(0,\sigma^2), n = 1,\dots,N,\ iid$ ו-$\mathbf{w}$ וקטור אקראי שמקיים: $w_j\sim N(0,\tau^2), j = 1,\dots,d,\ iid$ שהוא בת"ס ב- V_n. נרצה למצוא את משערך ה-MAP: נשתמש ב- (4) ובמקרה זה $\mathbf{x}$ הוא $\mathbf{w}$, נחשב את $p_{\mathbf{x}}(a)$ שהיא פונ' צפיפות הפילוג: $f_{\mathbf{w}}(\mathbf{w}) = \prod_{j=1}^d(2\pi\tau^2)^{-1/2}\exp\{\frac{1}{2\tau^2}(w_j)^2\}$ נרכיב log: $\log f_{\mathbf{w}}(\mathbf{w}) = \frac{-d}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\sum_{j=1}^d w_j^2 = -\frac{d}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\|\mathbf{w}\|^2\ (*)$. נחשב את $p_{y|x}(y\mid a)$, לפיכך צריך את התפלגות $y_n\mid\mathbf{w}$. בהינתן $\mathbf{w}$ נבחין ש-y_n יתפלג כמו V_n למעט הזזה בקבוע ($\mathbf{w}$) ולכן: $y_n\mid\mathbf{w}\sim N(w_0+\mathbf{x}_n^T\mathbf{w},\sigma^2)$ וצפיפות הפילוג: $f_{\mathbf{y}|\mathbf{w}} = \prod_{n=1}^N f_{y_n|\mathbf{w}}(y_n\mid\mathbf{w})$ כאשר: $f_{y_n|\mathbf{w}}(y_n\mid\mathbf{w}) = (2\pi\sigma^2)^{-1/2}\exp\{-\frac{1}{2\sigma^2}(y_n-w_0-\mathbf{x}_n^T\mathbf{w})^2\}$ הרכבת log: $\log f_{y_n|\mathbf{w}} = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n-w_0-\mathbf{x}_n^T\mathbf{w})^2$ ולכן: $\log f_{\mathbf{y}|\mathbf{w}} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^N(y_n-w_0-\mathbf{x}_n^T\mathbf{w})^2(**)$ ולכן: $\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}}\log f_{\mathbf{y}|\mathbf{w}}(\mathbf{y}\mid\mathbf{w}) + \log f_{\mathbf{w}}(\mathbf{w})$, נציב את $(*)$ ו-$(**)$ תוך התעלמות מגורמים שלא תלויים ב-$\mathbf{w}$, הכפלה ב-σ^2 והפיכת סימן ונקבל: $\hat{\mathbf{w}}_{MAP} = \arg\min_{\mathbf{w}}\left\{\sum_{n=1}^N(y_n-w_0-\mathbf{x}_n^T\mathbf{w})^2 + \frac{\sigma^2}{\tau^2}\|\mathbf{w}\|^2\right\}$ - אם נגדיר $\lambda = \frac{\sigma^2}{\tau^2}$ נקבל בדיוק את Ridge Regression למעט העובדה שהרעש כאן הוא אקראי (τ^2 מאוד קטן יניב λ מאוד גדולה ובכך יאלץ את $\mathbf{w}$ להיות בעל פרמטרים קטנים, σ^2 מאוד קטן יניב λ מאוד קטן). בצורה דומה עם פרמטרים זהים למעט: $w_j\sim Laplace(0,\tau^2), j = 1,\dots,d\ iid$ (תוחלת $\mu = 0$ ושונות $2b^2 = \tau^2$ ופונ' הפילוג: $(\frac{1}{2b}\exp(-\frac{|x-\mu|}{b}))$ ובת"ס ל V_n עם פיתוח דומה נוכל לקבל את LASSO.



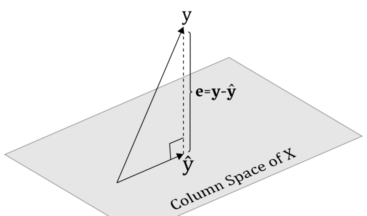
Hard-Margin SVM



Soft-Margin SVM



כדי לחשב את המרחק בין הנקודה x לבין העל-מישור L מגדירים את הוקטור x-x' ומטילים אותו על $\mathbf{w}$



השגיאה $\mathbf{e}$=$\mathbf{y}$-$\hat{\mathbf{y}}$ היא וקטור האורתוגונלי ל-$\hat{\mathbf{y}}$ ולכל וקטור אחר במרחב העמודות של X.

אלגברה לינ' - אם M מטריצה $n\times n$ סימטרית ממשית אז: (1) היא **$P.D$** אם $\mathbf{x}^TM\mathbf{x}>0$ לכל $\mathbf{x}\neq 0$ או שכל ע"ע חיוביים או שקיימת מטריצה B **הפיכה** המקיימת $M=B^TB$. (2) היא **$P.S.D$** אם $\mathbf{x}^TM\mathbf{x}\geq 0$ או שכל ע"ע אי-שליליים או שקיימת מטריצה B כך ש- $M=B^TB$. (3) אם M היא $P.D$ (או $P.S.D$) אז אברי האלכסון הם חיוביים (או אי-שליליים). בצורה סימטרית עבור $N.D$ ו- $N.S.D$. (4) ע"ע ממשיים ווקטורים עצמיים אורתוג' אחד לשני. (5) ניתנת **לפירוק ספקטרלי** - מטריצת ו"ע $=U$ (שהיא אורתוג' (העמודות שלה אורתוג')), מטריצת ע"ע $=\Lambda$ ואז: $M=U\Lambda U^T$ וגם $UU^T=U^TU=I$. אם M היא PSD אז $A+A^T$ היא PSD.

אם B **בסיס אורתונ'** אז כל הוקטורים אורתוג' ומנורמלים. לכל מטריצה A: מכפלת ע"ע $=\det(A)$ וגם: סכום ע"ע $=\mathrm{tr}(A)$. עקבה מקיימת: $\mathrm{tr}(AB)=\mathrm{tr}(BA);\mathrm{tr}(A)=\mathrm{tr}(A^T);\mathrm{tr}(BAB^{-1})=\mathrm{tr}(A);\mathrm{tr}(\mathbf{x}^TA\mathbf{x})=\mathrm{tr}(\mathbf{x}\mathbf{x}^TA);\mathrm{tr}(\alpha A)=\alpha\mathrm{tr}(A)$ טריק: $\sum_n\mathbf{x}_n^TA\mathbf{x}=\mathrm{tr}(A\sum_n\mathbf{x}_n\mathbf{x}_n^T)=\sum_n\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^TA)$. אם A לכסינה אז קיימת P הפיכה כך ש- $A=P^{-1}DP$ ו-$D=diag(\lambda_1,\dots,L_n)$ (λ_i ע"ע של A) וגם A דומה ל-D. אם $A_{n\times n}$ **דומה** ל-$B_{n\times n}$ אז: קיימת P הפיכה כך ש- $B=P^{-1}AP$ וגם: יש להן דרגה זהה, דטרמיננטה זהה, ע"ע זהים, פול' אופייני ומינימלי זהים ור"א/ר"ג זהים. כפל: $AB=\begin{bmatrix}[A]_1^rB\\ \vdots\\ [A]_m^rB\end{bmatrix}=\left[A[B]_1^c\dots A[B]_p^c\right]$. היפוך: $(AB)^{-1}=B^{-1}A^{-1}$ היפוך+שחלוף: $(A^T)^{-1}=(A^{-1})^T$. הגרדיאנט מוגדר בתור **וקטור עומד** $\frac{\partial f}{\partial x_n}(\mathbf{x})=\nabla f(\mathbf{x})$. ההסיאן מוגדרת בתור המטריצה $\frac{\partial f}{\partial x_i,x_j}=\nabla^2f(\mathbf{x})$ והיא **סימטרית** עבור פונקציה הגזירה פעמיים שנגזרתה הראשונה רציפה. לכל $\mathbf{w}$: $\nabla_{\mathbf{w}}(\mathbf{w}^T\mathbf{b})=\mathbf{b}$ ו- $\nabla_{\mathbf{w}}(\mathbf{w}^TA\mathbf{w})=(A+A^T)\mathbf{w}$. נגזרות מוכרות: $\nabla_{\mathbf{x}}(\mathbf{v}^T(A\mathbf{x}-\mathbf{b}))=A^T\mathbf{v}$, $\frac{\partial}{\partial X}\mathbf{a}^TX^T\mathbf{b}=\mathbf{b}\mathbf{a}^T$, $\frac{\partial}{\partial A}\mathbf{x}^TA\mathbf{x}=\frac{\partial}{\partial A}\mathrm{tr}(\mathbf{x}^T\mathbf{x}A)=(\mathbf{x}\mathbf{x}^T)^T=\mathbf{x}\mathbf{x}^T$. נוסחאות: $\mathbf{a}^TX^TX\mathbf{a}=(X\mathbf{a})^T(X\mathbf{a})=\|X\mathbf{a}\|^2$. אופטימום מקומי מקיים $\nabla f(\mathbf{x}^*)=\mathbf{0}$ ונקודה $\mathbf{x}^*$ היא **חשודה לקיצון.** אם $\mathbf{x}^*$ חשודה לקיצון אז: הנקודה $\mathbf{x}^*$ היא מינימום מקומי עבור $\nabla^2f(\mathbf{x}^*)\succ\mathbf{0}$ ומקסימום מקומי עבור $\nabla^2f(\mathbf{x}^*)\prec\mathbf{0}$ (אם זה לכל $\mathbf{x}$ אז זו נקודת אופטימום גלובלית). אם f גזירה פעמיים ברציפות ו $\nabla^2f(\mathbf{x}^*)\succeq 0$ אז $\mathbf{x}^*$ נק' מינ' גלובלית. פונקציה f תקרא **קונבקסית** אם לכל 2 נקודות $\mathbf{x},\mathbf{y}$ נוכל להגדיר קו ישר ℓ כך שלכל $\lambda\in[\mathbf{x},\mathbf{y}]$ מתקיים $\ell(\boldsymbol{\lambda})\geq f(\boldsymbol{\lambda})$. f קונבקסית $\iff\nabla^2f\succeq 0$. אם f קונבקסית וישנה נקודה $\mathbf{x}^*$ בה הנגזרת שלה מתאפסת - זוהי בהכרח **נקודת מינימום גלובלי.** פיתוח טיילור $f(\mathbf{y})=f(\mathbf{x})+\nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x})+\frac12(\mathbf{y}-\mathbf{x})^T\nabla^2f(\xi)(\mathbf{y}-\mathbf{x})$ כך ש- $\xi\in[\mathbf{x},\mathbf{y}]$. פונקציה אפינית: מהצורה $f=A\mathbf{x}+\mathbf{b}$. הטלה של $\mathbf{x}$ על $\mathbf{v}$: $\frac{\langle\mathbf{x},\mathbf{v}\rangle}{\langle\mathbf{v},\mathbf{v}\rangle}\mathbf{v}$

הוכחות: (1) X דרגה מלאה $\iff X^TX$ הפיכה: $\Leftarrow$ נניח בשלילה שלא, אז קיים $\mathbf{c}\neq\mathbf{0}$ כך ש- $X^TX\mathbf{c}=0$, נכפיל ב-$\mathbf{c}$ ונקבל: $\mathbf{c}^TX^TX\mathbf{c}=0$ ולכן $\|X\mathbf{c}\|^2=0\Rightarrow X\mathbf{c}=0$ וזו סתירה לכך ש X דרגה מלאה. $\Rightarrow$ נניח בשלילה ש X^TX פיכה וש-X לא דרגה מלאה, אז קיים $\mathbf{c}\neq\mathbf{0}\in\ker(X)$ ולכן $X\mathbf{c}=0$ ולכן גם $X^TX\mathbf{c}=0$ לפיכך $\mathbf{c}$ ו"ע של X^TX ולכן לא הפיכה - סתירה. (2) אם $J_i(\{r_{nk}\}^i,\{\boldsymbol{\mu}_k\}^i)$ הוא מינ' גלובלי לבעית ה-K-means עבור i נקודות ו- $J_{i+1}(\{r_{nk}\}^{i+1},\{\boldsymbol{\mu}_k\}^{i+1})$ מינ' גלובלי עבור $i+1$ נק' (אותן i בתוספת נקודה חדשה) אז $J_{i+1}(\{r_{nk}\}^{i+1},\{\boldsymbol{\mu}_k\}^{i+1})\geq J_i(\{r_{nk}\}^i,\{\boldsymbol{\mu}_k\}^i)$: נניח בשלילה שלא, אז: $J_{i+1}(\{r_{nk}\}^{i+1},\{\boldsymbol{\mu}_k\}^{i+1})<J_i(\{r_{nk}\}^i,\{\boldsymbol{\mu}_k\}^i)$ ולכן אם נחסיר את הנקודה $i+1$ מ-J_{i+1} ערך הפתרון שלו ישאר זהה או ייקטן (כי החסרנו ערך אי-שלילי), משמע: $J_i(\{r_{nk}\}^i,\{\boldsymbol{\mu}_k\}^i)>J_{i+1}(\{r_{nk}\}^{i+1},\{\boldsymbol{\mu}_k\}^{i+1})\geq J_i(\{r_{nk}\}^{i+1},\{\boldsymbol{\mu}_k\}^{i+1})$ וזה סותר את הטענה ש- $J_i=\{r_{nk}\}^i,\{\boldsymbol{\mu}_k\}^i$ הוא מינ' גלובלי.

הסתברות וסטט': השונות: $\sigma^2\triangleq E[(X-\mu)^2]=E[X^2]-E[X]^2$ מקיימת $\mathrm{Var}(aX+b)=\mathrm{Var}(aX)=a^2\mathrm{Var}(X)$. עבור X_1,X_2 בת"ס (או חסרי קורלציה): $\mathrm{Var}(X_1+X_2)=\mathrm{Var}(X_1)+\mathrm{Var}(X_2)$ עבור מטריצה A ווקטור $\mathbf{y}$: $\mathrm{Var}(A\mathbf{y})=A\mathrm{Var}(\mathbf{y})A^T$. התוחלת: מסומנת $E[X]$ או μ. קורלציה: מוגדרת בתור $r_{XY}=E[XY]$, מטריצה: $X^TX=\sum_i\mathbf{x}_i\mathbf{x}_i^T$. מקדם קורלציה בין X ל-Y: $\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$ ו- $|\rho_{XY}|\leq 1$. שונות משותפת: $\sigma_{XY}=E[(X-\mu_X)(Y-\mu_Y)]$ או $\sigma_{XY}=r_{XY}-\mu_X\mu_Y$. המטריצה תמיד סימטרית והיא P.S.D. בייס: $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}\mid\mathbf{y})=\frac{p_{\mathbf{y},\mathbf{x}}(\mathbf{y},\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})}=\frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})}$. צפיפות פילוג של מ"א גאוסי $X\sim N(\mu,\sigma^2)$: $f_X(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$, במקרה הוקטורי עבור מ"א $\mathbf{X}\in\mathbb{R}^N$, צריך שמטריצת ה-cov ($C_{\mathbf{XX}}$) תהיה P.D (כדי להיות הפיכה) ואז: $f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\mu},C_{\mathbf{XX}})=\frac{1}{\sqrt{(2\pi)^N\det(C_{\mathbf{XX}})}}\exp\{-\frac12(\mathbf{X}-\boldsymbol{\mu})^TC_{\mathbf{XX}}^{-1}(\mathbf{X}-\boldsymbol{\mu})\}$

נוסחאות ומינוחים: הבעיה הפרימלית: בהנתן בעית אופטימיזציה $\min_{\mathbf{x}\in\mathbb{R}^n}f_0(\mathbf{x}),\ \text{s.t}\ \forall i: f_i(\mathbf{x})\leq 0,\forall j: h_j(\mathbf{x})=0$ כאשר תחום ההגדרה לא ריק - נסמן את הפתרון האופטימלי לבעיה: $f_0(\mathbf{x})=p^*$ ולוקטור $\mathbf{x}$ המתאים לו נקרא primal optimal. בעית QP: מהצורה $\min_{u\in R^L}\frac12\mathbf{u}^TQ\mathbf{u}+\mathbf{p}^T\mathbf{u}\ \text{s.t}\ A\mathbf{u}\geq\mathbf{c}$ עבור $Q_{L\times L},\mathbf{p}_{L\times1},A_{M\times L},\mathbf{c}_{M\times1}$ (**ואם** Q היא P.S.D אז הבעיה היא **Convex QP**) שהפתרון הוא $\mathbf{u}^*\leftarrow\mathrm{QP}(Q,\mathbf{p},A,\mathbf{c})$. הלגרנג'יאן: $\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu})=f_0(\mathbf{x})+\sum_{i=1}^m\lambda_if_i(\mathbf{x})+\sum_{j=1}^p\nu_jh_j(\mathbf{x})$. הפונקציה הדואלית: עבור $\boldsymbol{\lambda}\in\mathbb{R}^m,\boldsymbol{\nu}\in\mathbb{R}^p$ נחפש $\mathbf{x}$ בתחום ההגדרה D כך ש: $g(\boldsymbol{\lambda},\boldsymbol{\nu})=\inf_{\mathbf{x}\in D}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu})=\inf_{\mathbf{x}\in D}\left(f_0(\mathbf{x})+\sum_{i=1}^m\lambda_if_i(\mathbf{x})+\sum_{j=1}^p\nu_jh_j(\mathbf{x})\right)$ וזו מקיימת שלכל $\boldsymbol{\lambda}\geq\mathbf{0}$ ולכל $\boldsymbol{\nu}$: $g(\boldsymbol{\lambda},\boldsymbol{\nu})\leq p^*$ כי לכל $\tilde{\mathbf{x}}$ נק' פיסיבילית (ובפרט אופטימלית): $g(\boldsymbol{\lambda},\boldsymbol{\nu})=\inf_{\mathbf{x}\in D}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu})\leq\mathcal{L}(\tilde{\mathbf{x}},\boldsymbol{\lambda},\boldsymbol{\nu})\leq f_0(\tilde{\mathbf{x}})\leq p^*$. אם $\boldsymbol{\lambda}\geq 0$ ו- $(\boldsymbol{\lambda},\boldsymbol{\nu})\in D$ אז $(\boldsymbol{\lambda},\boldsymbol{\nu})$ נקראים dual feasible. הבעיה הדואלית: **תמיד קונבקסית** ומוגדרת $\max g(\boldsymbol{\lambda},\boldsymbol{\nu})\ \text{s.t}\ \boldsymbol{\lambda}\geq 0$ והפתרון שלה: d^* שמורכב מ- $(\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$ נקרא dual optimal. duality-gap: מוגדר p^*-d^* והוא תמיד אי-שלילי. דואליות חלשה: מוגדרת בתור $d^*\leq p^*$. דואליות חזקה: מוגדרת בתור $d^*=p^*$. בעית אופט' קונבקסית: $\min f_0(\mathbf{x})\ \text{s.t}\ \forall i: f_i(\mathbf{x})\leq 0, A\mathbf{x}=\mathbf{b}$ ו- f_i קונבקסית לכל i. תאורית Slater: אם בעית האופט' קונבקסית וקיים $\mathbf{x}$ שמקיים את תנאי האי-שוויון בצורה חזקה ($f_i(\mathbf{x})<0$) אז יש דואליות חזקה. הרחבה ל-Slater: מספיק שיהיו k פונקציות שתנאי האי -שווין חזק מתקיים **ועוד** $m-k$ פונקציות אפיניות עם אי-שיוון חלש. Complementary Slackness: אם $\mathbf{x}^*$ הוא primal optimal ו- $(\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$ הם dual optimal ומתקיימת דואליות חזקה אז: (1) $\mathbf{x}^*$ מביא למינימום את הלגרנג'יאן. (2) $\sum_{i=1}^m\lambda_if_i(\mathbf{x}^*)=0$ ולכן $\lambda_if_i(\mathbf{x}^*)=0$. (3) $\lambda_i^*>0\Rightarrow f_i(\mathbf{x}^*)=0$ **או** $f_i(\mathbf{x}^*)<0\Rightarrow\lambda_i^*=0$. תנאי KKT: אם $\mathbf{x}^*$ הוא primal optimal ו- $(\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$ הם dual optimal ומתקיימת דואליות חזקה **וגם** $f_0,f_1,\dots f_m,h_1,\dots h_p$ גזירות: כיוון ש $\mathbf{x}^*$ מביא למינימום את $\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$ מעל $\mathbf{x}$ אז $\nabla\mathcal{L}$ מתאפס ב- $\mathbf{x}^*$, והתנאים הבאים מתקיימים: (1) $f_i(\mathbf{x}^*)\leq 0,i=1,\dots,m$ (2) $h_i(\mathbf{x}^*)=0,i=1,\dots,p$ (3) $\lambda_i^*\geq 0,i=1,\dots,m$ (4) $\lambda_i^*f_i(\mathbf{x}^*)=0,i=1,\dots,m$ (5) $\nabla\mathcal{L}(\mathbf{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)=0$. תנאי KKT עבור בעיה קונבקסית: אם הבעיה הפרימלית קונבקסית וישנם $\tilde{\mathbf{x}},\tilde{\boldsymbol{\lambda}},\tilde{\nu}$ שמקיימים את תנאי ה-KKT אז הם מהווים תנאי מספיק לדואליות חזקה (הם פתרונות לבעיה הפרימלית והדואלית ויש דואליות חזקה).

קרנלים: עבור K_1,K_2 קרנלי ולידיים גם הקרנלים הבאים ולידיים: $K=K_1+K_2$, $K=\lambda K_1$, $K=K_1K_2$ כמו-כן $K(\mathbf{x},\mathbf{y})=\exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$: RBF ולידי וגם $K(\mathbf{x},\mathbf{y})=(\mathbf{x}^T\mathbf{y}+c)^q$ ולידי עבור $c>0$ לכל $q\in\mathbb{N}$.

ממנים:

11: **1.** אם המידע פריד אז פרספטרון תמיד מתכנס ל-$\mathbf{w}^*$ ומס' האיטרציות חסום: נגדיר $R=\max_n\|\mathbf{x}_n\|$ ו- $\rho=\min_ny_n(\mathbf{w}^{*T}\mathbf{x}_n)>0$ ואז: $t\leq\frac{R^2\|\mathbf{w}^*\|^2}{\rho^2}$. **3.** X היא מטריצה ששורותיה הן $\mathbf{x}_n^T$, מתקיים: $\mathcal{N}(X^T)=\mathcal{R}(X)^\perp;\mathcal{N}(X)=\mathcal{R}(X^T)^\perp$ ו- $\mathcal{R}(X^T)=\mathcal{N}(X)^\perp;\mathcal{R}(X)=\mathcal{N}(X^T)^\perp$ כמו-כן: $\mathcal{R}(X^TX)=\mathcal{R}(X^T)$. כאשר X לא מדרגה מלאה למשוות הנורמליות יש יותר מפתרון אחד ולכל 2 פתרונות $\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2$ מתקיים: $X(\hat{\mathbf{w}}_1-\hat{\mathbf{w}}_2)=\mathbf{0}$ (ההפרש בקרנל של X). ההטלה של $\mathbf{y}$ ל- $\mathcal{R}(X)$ יחידה ומוגדרת $\hat{\mathbf{y}}=X\hat{\mathbf{w}}$ כאשר $\hat{\mathbf{w}}$ פתרון למשוואות הנורמליות. **4.** נגדיר $H=X(X^TX)^{-1}X^T$, כל הע"ע של H הם 0 או 1 ומס' הע"ע שהם 1 הוא $d+1$ ודרגת H היא $d+1$. **5.** אם משתמשים בשגיאה ריבועית על רגרסיה לינ' אז ההיפותזה מתכנסת ל- $\frac1N\sum_{n=1}^Ny_n$ (מינ' גלובלי ע"פ נגזרת שניה) ואם משתמשים בשגיאת ערך מוחלט מתכנסים ל- h_{med} (חציון).

12: **1.** אם α^* ממקסם את הבעיה הדואלית של SVM ו-Δ מרחק מינימלי בין נקודה למפריד אז $L(\boldsymbol{\alpha}^*)=\frac12\sum_{n=1}^N\alpha_n^*$ לכן $\|\mathbf{w}\|^2=\sum_{n=1}^N\alpha_n^*$ וגם $\Delta=(\sqrt{\sum_{n=1}^N\alpha_n^*})^{-1}$. **2.** אם המידע פריד לינ' אז יש לכל היותר $d+1$ essential SVs. **4.** אם P פתרון ל- Soft SVM אז בהנחה ש $P\to\infty$ כאשר $C\to\infty$ בהכרח המידע לא פריד כי אם היה פריד $\xi_n\to 0$ ולכן בהכרח $P\nrightarrow\infty$ כאשר $C\to\infty$. אם P פתרון אופט' ל-Soft SVM והמידע פריד ונסמן את Δ_0 בתור המרחק של המפריד האופט' מהנקודה הקרובה ביותר אליו אז: $\Delta_0\leq 1/\sqrt{2P}$ כי נתכנס לפתרון של Hard, נסמנו P_H ונקבל: $P_H=\frac12\|\mathbf{w}^*\|$ ולכן $\Delta_H=1/\|\mathbf{w}^*\|=1/\sqrt{2P_H}$ ומכך ש- P_H אופטימלי בהכרח: $\Delta_0\leq\Delta_H\Rightarrow\Delta_0\leq 1/\sqrt{2P}$. יכול להתקיים $\alpha_n^*=0$ וגם $(\mathbf{x}_n,y_n)$ על שולי המפריד ומקיים שוויון באילוץ - $\mathbf{x}_1=[0,0]^T,\mathbf{x}_2=[1,0]^T,\mathbf{x}_3=[0,1]^T$ ותיוגים: $y_1=+1,y_2=+1,y_3=-1$, נפתור את האילוצים ונקבל $w_1^*=0,w_2^*=-2$ ו- $b^*=1$ ו- $\mathbf{w}^*=\alpha_1^*\mathbf{x}_1+\alpha_2^*\mathbf{x}_2-\alpha_3^*\mathbf{x}_3=[\alpha_2^*,-\alpha_3^*]^T$ נסיק: $[[\alpha_2^*,-\alpha_3^*]^T=[w_1^*,w_2^*]=[0,2]$ ולכן $\alpha_2^*=0$ אבל מקיים את האילוץ בשוויון ולכן SV. אם $\alpha_n^*=0$ אז $\mathbf{x}_n$ הוא לא essential ולכן אפשרי לזרוק אותו (לא ישפיע).

14: **5.** השגיאה של K-means מונו' יורדת ולכן האלגוריתם חייב להתכנס.

דוגמאות: SVM עם משקל v_n לכל $\mathbf{x}_n$: הבעיה: $\min_{\mathbf{w},b,\boldsymbol{\xi}}\frac12\|\mathbf{w}\|^2+C\sum_{n=1}^Nv_n\xi_n\ \text{s.t}\ y_n(\mathbf{w}^T\mathbf{x}_n+b)\geq 1-\xi_n,\xi_n\geq 0$ הלגרנגיאן: $\mathcal{L}(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{r})=\frac12\mathbf{w}^T\mathbf{w}+C\sum_{n=1}^Nv_n\xi_n-\sum_{n=1}^N\alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n+b)-1+\xi_n)-\sum_{n=1}^Nr_n\xi_n$ גזירות: $\frac{\partial}{\partial\mathbf{w}}=\mathbf{w}-\sum_{n=1}^N\alpha_ny_n\mathbf{x}_n=0$, $\frac{\partial}{\partial b}=-\sum_{n=1}^N\alpha_ny_n=0$, $\frac{\partial}{\partial\xi_n}=Cv_n-\alpha_n-r_n=0$ ולכן: $\mathbf{w}=\sum_{n=1}^N\alpha_ny_n\mathbf{x}_n$ ו- $\sum_{n=1}^N\alpha_ny_n=0$. הצבה ב-$\mathcal{L}$: $\mathcal{L}(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{r})=\frac12\mathbf{w}^T\mathbf{w}+C\sum_{n=1}^Nv_n\xi_n-\mathbf{w}^T(\sum_{n=1}^N\alpha_ny_n\mathbf{x}_n)-b(\sum_{n=1}^N\alpha_ny_n)+\sum_{n=1}^N\alpha_n-\sum_{n=1}^N\xi_n\alpha_n-\sum_{n=1}^Nr_n\xi_n$ לאחר צמצום נקבל את הבעיה הדואלית המתאימה: $\max_{\boldsymbol{\alpha}}-\frac12\sum_{n=1}^N\sum_{m=1}^N\alpha_n\alpha_my_ny_m\mathbf{x}_n^T\mathbf{x}_m+\sum_{n=1}^N\alpha_n\ \text{s.t}\ r_n\geq 0,\alpha_n\geq 0$.

SVM עם $C/2$: הבעיה: $\min\frac12\mathbf{w}^T\mathbf{w}+\frac{C}{2}\sum_{n=1}^N\xi_n^2\ \text{s.t}\ y_n(\mathbf{w}^T\mathbf{x}_n+b)\geq 1-\xi_n$, לגרנג'יאן: $\mathcal{L}(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{r})=\frac12\mathbf{w}^T\mathbf{w}+\frac{C}{2}\sum_{n=1}^N\xi_n^2-\sum_{n=1}^N\alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n+b)-1+\xi_n)$ גזירות: $\frac{\partial}{\partial\mathbf{w}}=\mathbf{w}-\sum_{n=1}^N\alpha_ny_n\mathbf{x}_n=0$, $\frac{\partial}{\partial b}=-\sum_{n=1}^N\alpha_ny_n=0$ ו- $\frac{\partial}{\partial\xi_n}=C\xi_n-\alpha_n=0$ ולכן: $\mathbf{w}=\sum_{n=1}^N\alpha_ny_n\mathbf{x}_n$, $\xi_n=\frac{\alpha_n}{C}$ ו- $\sum_{n=1}^N\alpha_ny_n=0$ הצבה ב-$\mathcal{L}$: $\mathcal{L}(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{r})=-\frac12\sum_{n=1}^N\sum_{m=1}^N\alpha_n\alpha_my_ny_m\mathbf{x}_n^T\mathbf{x}_m+\sum_{n=1}^N\alpha_n-\frac12\sum_{n=1}^N\frac{\alpha_i^2}{C}$ הבעיה הדואלית: $\max_\alpha\sum_{n=1}^N-\frac12\sum_{n=1}^N\sum_{m=1}^N\alpha_n\alpha_my_ny_m\mathbf{x}_n^T\mathbf{x}_m+\sum_{n=1}^N\alpha_n-\frac12\sum_{n=1}^N\frac{\alpha_n^2}{C}\ \text{s.t}\ \alpha_n\geq 0,\sum_{n=1}^N\alpha_ny_n=0$.

כללי: **האם קיים מקרה בו יש פתרון ל-Soft SVM שאין נקודות על השוליים ועם $\mathbf{w}\neq 0$?** כן, עבור $x_1=1,y_1=1,x_2=-1,y_2=-1$ ו- $C=1/8$ נקבל $w=1/4$ ושתי הנקודות x_1,x_2 בתוך שולי המפריד והאלפות $\alpha_1=\alpha_2=C$.

שימוש ב-PCA: עבור $\mathcal{D}=\{(-2,10),(-1,5),(0,0),(1,5),(2,10)\}$ נמצא את PC1, תחילה $X^T=\begin{bmatrix}-2&-1&0&1&2\\10&5&0&5&10\end{bmatrix}$, נחשב ממוצע של כל שורה: $\overline{r_1}=0,\overline{r_2}=6$ ונחסיר מהשורה: $X_0^T=\begin{bmatrix}-2&-1&0&1&2\\4&-1&-6&-1&4\end{bmatrix}$ מטריצת ה-cov: $X_0^TX_0=\begin{bmatrix}0&10\\70&0\end{bmatrix}$ וע"ע הגדול ביותר $\lambda_1=70$ ו $\text{PCA=}\mathbf{u}_1=\begin{bmatrix}1\\0\end{bmatrix}$

מרחק בין על-מישור לנקודה: $d(\mathbf{x}_n,\mathbf{w},b)=\frac{|\mathbf{w}^T\mathbf{x}_n+b|}{\|\mathbf{w}\|}$. **תיקון על-מישור $\mathbf{w}^{*T}\mathbf{x}+b$ על פי נקודה** $\mathbf{x}_n$: נסמן $\Delta^*=1/\|\mathbf{w}\|$ ו- $\tilde{\Delta}=d(\mathbf{x}_n,\mathbf{w}^*,b^*)$. אם $\mathbf{x}_n$ סווגה נכון אז: $\Delta=\Delta^*+\tilde{\Delta}$ אחרת: $\Delta=\Delta^*-\tilde{\Delta}$ ואז: $1/\|\mathbf{w}\|=\frac{\Delta}{2}=|\alpha|\Delta^*$ נמצא את α ונקבל את $\mathbf{w}=\frac{\mathbf{w}^*}{\alpha}$. נמצא את b: $\mathbf{w}^T\mathbf{x}_n+b=y_n$ ונקבל את $g(\mathbf{x})=sign(\mathbf{w}^T\mathbf{x}+b)$.

בדיקת שוני בין חזאים: בהנתן $\hat{\mathbf{y}}_1,\hat{\mathbf{y}}_2$ ותיוגים $\mathbf{y}$ נבדוק: $\|\hat{\mathbf{y}}_2-\mathbf{y}\|^2=\|\hat{\mathbf{y}}_2-\mathbf{y}_1+\mathbf{y}_1-\mathbf{y}\|^2=\|\hat{\mathbf{y}}_2-\hat{\mathbf{y}}_1\|^2+2(\hat{\mathbf{y}}_1-\mathbf{y})^T(\hat{\mathbf{y}}_2-\hat{\mathbf{y}}_1)+$ $\|\hat{\mathbf{y}}_2-\hat{\mathbf{y}}_1\|^2$ נשתמש באורתוג' של השגיאות (אם יש), נצמצם וננסה להגיע לאי שוויון עם $\|\hat{\mathbf{y}}_1-\mathbf{y}\|^2$.

רשתות נוירונים:

שם	סימון	סוג
סיגנל כניסה	$\mathbf{s}^{(\ell)}$	וקטור ממםד $d^{(\ell)}$
יציאה	$\mathbf{x}^{(\ell)}$	וקטור ממםד $d^{(\ell)}+1$
משקלי כניסה	$W^{(\ell)}$	מטריצה ממםד $(d^{(\ell-1)}+1)\times d^{(\ell)}$
משקלי יציאה	$W^{(\ell+1)}$	מטריצה ממםד $(d^{(\ell)}+1)\times d^{(\ell+1)}$

לא לשכוח

נגזרות:

- $ln(f(x))' = \frac{1}{x} \cdot f'(x)$
- כלל השרשרת: $h'(x) = f'(g(x))g'(x)$,
- מכפלה $h'(x) = f'(x)g(x) + g'(x)f(x)$
- חלוקה: $(f(x)/g(x))' = \frac{f'(x)g(x) - g'(x)f(x)}{g(x)^2}$.

$\text{tr}(ABCD) = \text{tr}(BCDA) = \text{tr}(CDAB) = \text{tr}(DABC)$ ועבור A, B ריבועיות: $\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$

מטריצות קרוס-קווריאנס:

$C_{xx} = \frac{1}{N}\sum_{n=1}^{N}(x_i - \overline{x})^2; C_{xy} = \frac{1}{N}\sum_{n=1}^{N}(x_i - \overline{x})(y_i - \overline{y}); C_{yy} = \frac{1}{N}\sum_{n=1}^{N}(y - \overline{y})^2$