מבוא ללמידה חישובית (20942) | סיכום הרצאה 6

מנחה: שי מימון
סמסטר: 2022א'
נכתב על ידי: מתן כהן

1 חזרה המשך - Hard-Margin SVM - The separable case

בשיעור הקודם ניסחנו את הבעיה כבעית אופטימיזציה בה רצינו למצוא את המרחק המינימלי מבין מרחקי הנקודות מסט הדוגמאות אל העל-מישור (Hyperplane) בעזרת:

$$d(\mathbf{x}_n, \mathbf{w}, b) = \frac{\left|\mathbf{w}^T\mathbf{x} + b\right|}{\|\mathbf{w}\|} \overset{\text{separable}}{=} \frac{y_n(\mathbf{w}^T\mathbf{x}_n + b)}{\|\mathbf{w}\|}$$

ואת המרחק המינימלי הנ"ל נרצה למקסם לפי $\mathbf{w}$ ו-b - בעצם למצוא את אותו על-מישור שימקסם את המרחק המינימלי שמצאנו.

את בעית האופטימיזציה הנ"ל רצינו לפתור בעת שהגדרנו את האילוץ שהעל-מישור מפריד באופן מושלם את הדוגמאות:

$$\text{to subject } y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right) > 0 \;\; n = 1, .., N$$

לאחר מכן הבנו שכל עוד יש לנו על-מישור מסוים, לא משנה כיצד נעשה סקיילינג (scale) בעזרת ערך חיובי כלשהו ρ ל- $(\mathbf{w}, b)$ (ביחד כמובן), לא נשנה את העל-מישור ולא את איזורי ההחלטה.

בחרנו את $(\mathbf{w}, b)$ כך ש:

$$\min_{n=1,..,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$$

בעזרת הדרישה הזו הצלחנו לפשט את בעית האופטימיזציה כאשר:

$$\min_{n=1,..,N} d(\mathbf{x}_n, \mathbf{w}, b) = \frac{\min_{n=1,..,N} y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

לבסוף הגענו לבעית האופטימיזציה הסופית:

$$\underset{\mathbf{w},b}{\text{minimize}} \; \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{subject to } y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1, \;\; n = 1, .., N$$

הערה: ניתן להבחין שפונקצית המטרה (שרוצים להביא למינימום) היא קוואדרטית והאילוצים הם לינאריים.

בעיה זו שייכת למשפחת הבעיות הנקראות quadratic programming להן יש פתרונות יעילים.

מה שאנחנו צריכים לדעת לעשות זה לקחת בעיה כלשהי ולהמירה לצורה סטנדרטית של quadratic programmin.

2 Quadratic Programming (QP)

2.1 רקע

ב - QP יש בעיית מינימיזציה קונבקסית אשר בהינתן

(1) **מטריצה** $Q_{L\times L}$ שהיא חיובית למחצה (PSD)

(2) **וקטור** $\mathbf{p}_{L\times 1}$

(3) **מטריצה** $A_{M\times L}$ שמוגדרת כ: $A = \begin{bmatrix} -\mathbf{a}_1^T- \\ \vdots \\ -\mathbf{a}_M^T- \end{bmatrix}$

(4) **וקטור** של אילוצים $\mathbf{c}_{M\times 1}$

נוכל למצוא את הפתרון $\mathbf{u}^* \leftarrow \text{QP}(Q,\mathbf{p},A,\mathbf{c})$ לבעיה:

$$\begin{aligned} &\underset{\mathbf{u}\in\mathbb{R}^L}{min}\ \frac{1}{2}\mathbf{u}^TQ\mathbf{u}+\mathbf{p}^T\mathbf{u} \\ &\text{subject to } A\mathbf{u}\succeq\mathbf{c} \end{aligned}$$

או בפירוק לוקטורים:

$$\begin{aligned} &\underset{\mathbf{u}\in\mathbb{R}^L}{min}\ \frac{1}{2}\mathbf{u}^TQ\mathbf{u}+\mathbf{p}^T\mathbf{u} \\ &\text{subject to } \mathbf{a}_m^T\mathbf{u}\geq c_m \ \ m=1,..,M \end{aligned}$$

2.2 פתרון בעית האפוטימיזציה של Hard-Margin SVM בעזרת QP

נרצה לבטא את הבעיה שלנו כבעיית QP:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}}\ \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &\text{s.t}\quad y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)\geq 1,\ \ n=1,..,N \end{aligned} \quad\Rightarrow\quad \begin{aligned} &\underset{\mathbf{u}\in\mathbb{R}^L}{min}\ \frac{1}{2}\mathbf{u}^TQ\mathbf{u}+\mathbf{p}^T\mathbf{u} \\ &\text{s.t}\ \ \mathbf{a}_m^T\mathbf{u}\geq c_m\ \ m=1,..,M \end{aligned}$$

נסמן: $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$ ונכתוב:

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}\begin{pmatrix} b & \mathbf{w}^T \end{pmatrix}\cdot Q\cdot\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$$

נגדיר את $Q_{(d+1)\times(d+1)}$ להיות מטריצת בלוקים שהבלוק השמאלי העליון הוא הסקלר 0 והימני התחתון הוא $I_{d\times d}$, בשאר המקומות נשתמש בוקטור ה-0 על מנת לשמור על הצורה הרצויה:

$$\frac{1}{2}\begin{pmatrix} b & \mathbf{w}^T \end{pmatrix}\cdot Q\cdot\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} = \underbrace{\frac{1}{2}\begin{bmatrix} b & \mathbf{w}^T \end{bmatrix}\cdot\begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & I_{d\times d} \end{bmatrix}\cdot\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}}_{\frac{1}{2}\mathbf{w}^T\mathbf{w}}$$

כיוון שהצלחנו להביע את המשוואה בצורה קוואדרטית, אין צורך להשתמש בתוספת הלינארית ולכן נגדיר $\mathbf{p}=\mathbf{0}$.

נתבונן כעת בחלק של האילוצים של ה-QP ונשתמש בהגדרת $\mathbf{u}=\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$ ממקודם.

נתחיל מלרשום אילוץ בודד:

$$y_n\left(\mathbf{w}^T\mathbf{x}_n+b\right)=y_n\cdot\begin{bmatrix}1 & \mathbf{x}_n^T\end{bmatrix}\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}=\underbrace{y_n\cdot\begin{bmatrix}1 & \mathbf{x}_n^T\end{bmatrix}}_{\mathbf{a}_n^T}\cdot\mathbf{u}\geq 1,\quad n=1,..,N$$

ובעצם המטריצה A שקיבלנו היא המטריצה X שהגדרנו בשיעורים קודמים שמרופדת ב-1 -ים:

$$A_{N\times(d+1)}=\begin{bmatrix} y_1\cdot\begin{bmatrix}1 & \mathbf{x}_1^T\end{bmatrix} \\ \vdots \\ y_N\cdot\begin{bmatrix}1 & \mathbf{x}_N^T\end{bmatrix} \end{bmatrix}=\underbrace{\begin{bmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_n \end{bmatrix}}_{Y_{N\times N}}\cdot X=Y\cdot X$$

נותר רק למצוא את וקטור $\mathbf{c}$ שעל פי האילוצים בבעיה שלנו הוא בעצם וקטור של אחדות $\mathbf{c}=\mathbf{1}$

כעת נרצה לוודא שהבעיה היא בעיה קונבקסית על ידי הוכחה ש Q היא PSD, על פי הגדרה:

$$\alpha^TQ\alpha=\sum_{i=2}^{d+1}\alpha_i^2\geq 0$$

לכן המטריצה חיובית למחצה ולפיכך גם קונבקסית - כל מינימום לוקאלי הוא גם גלובלי.

מצאנו $Q,\mathbf{p},A,\mathbf{c}$ וכעת נוכל להשתמש בהם ולקבל $\mathbf{u}^*=\begin{bmatrix} b^* \\ \mathbf{w}^* \end{bmatrix}$ שמגדיר את החזאי האופטימלי שלנו:

$$\boxed{g(\mathbf{x})=sign(\mathbf{w}^T\mathbf{x}+b)}$$

דוגמה: מהשיעור הקודם: $\mathbf{y}=\begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$, $X=\begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix}$, נגדיר:

$$Q_{3\times x}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\quad \mathbf{p}=\mathbf{0},\quad \mathbf{c}=\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},\quad A=\begin{bmatrix} -1 & 0 & 0 \\ -1 & -2 & -2 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \end{bmatrix}$$

וכאן נגמר הפתרון כאשר:

$$\mathbf{u}^*=\begin{bmatrix} b^* \\ \mathbf{w}^* \end{bmatrix}=QP(Q,\mathbf{p},A,\mathbf{c})\overset{\text{run QP solver}}{=}\begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$$

הערה: גודל ה-margin של העל-מישור (המפריד) הוא:

$$\frac{1}{\|\mathbf{w}^*\|}=\frac{1}{\left\|\begin{bmatrix} -1 & 1 & -1 \end{bmatrix}\right\|}=\frac{1}{\sqrt{2}}$$

3 דואליות - Duality

3.1 הלגראנז'יאן - $\mathcal{L}$agrangian והבעיה הפרימלית

נתבונן בבעית האופטימיזציה הבאה (הבעיה הפרימלית):

$$
\begin{array}{lll}
\textbf{minimize} & f_0(\mathbf{x}) & | \\
\textbf{s.t} & f_i(\mathbf{x}) \leq 0, \ \ i = 1,..,m & | \quad \mathbf{x} \in \mathbb{R}^n \\
 & h_i(\mathbf{x}) = 0, \ \ i = 1,..,p & | \\
 & \mathcal{D} = \left(\bigcap_{i=0}^{m} \mathrm{dom} f_i\right) \cap \left(\bigcap_{i=1}^{p} \mathrm{dom} h_i\right) \text{ - non empty} & |
\end{array}
$$

נסמן את הפתרון שלה ב- $\mathbf{p}^*$ (כיוון שבהמשך נקרא לה הבעיה הפרימלית - Primal Optimization Problem)

הגדרה: הלנגראנז'יאן

$$
\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \quad \mathrm{dom} L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p
$$

הסבר: $\mathbf{x}$ הוא הוקטור מבעית האופטימיזציה, הוקטורים $\boldsymbol{\lambda}$ ו- $\boldsymbol{\nu}$ הם **כופלי הלגראנז'** (משתנים דואלים) שלנו. הלגראנז'יאן מוסיף ל f_0 את הקובינציות הלינאריות שכוללות את כופלי לגראנז' והאילוצים

3.2 הפונקציה הדואלית - The Lagrange dual function

הגדרה: בעזרת ההגדרות שכבר הגדרנו נגדיר את **הפונקציה הדואלית**:

$$
g : \mathbb{R}^m \times \mathbb{R}^p \longrightarrow \mathbb{R}
$$

$$
g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \right), \quad \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\nu} \in \mathbb{R}^p
$$

טענה: לכל $\boldsymbol{\lambda} \geq 0$ ולכל $\boldsymbol{\nu}$: $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathbf{p}^*$

הוכחה. יהיו $\tilde{\mathbf{x}}$ ערך התואם את אילוצי בעית האופטימיזציה ו $\boldsymbol{\lambda} \geq 0$.

- מהגדרת הלגראנז'יאן:

$$
\mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{\mathbf{x}})
$$

- מהאילוץ $f_i(\mathbf{x}) \leq 0, \ \ i = 1,..,m$ נסיק:

$$
\sum_{i=1}^{m} \lambda_i f_i(\tilde{\mathbf{x}}) \leq 0
$$

- מהאילוץ $h_i(\mathbf{x}) = 0, \quad i = 1, .., p$ נסיק:

$$\sum_{i=1}^{p} \nu_i h_i(\tilde{\mathbf{x}}) = 0$$

- לכן:

$$\mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\tilde{\mathbf{x}})$$

- כמו-כן לכל $\tilde{\mathbf{x}}$ המקיים את אילוצי בעית האופטימיזציה ובפרט עבור הפתרון $\mathbf{p}^*$:

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x}) \\ &\Rightarrow g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathbf{p}^* \end{aligned}$$

■

הערה: כאשר $\lambda \geq 0$ ו- $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \text{dom } g$ נקרא לזוג המשתנים הללו **dual fessible**

דוגמה: ריבועים פחותים:

$$\begin{aligned} &\underset{\mathbf{x} \in \mathbb{R}^n}{min} \ \mathbf{x}^T \mathbf{x} \\ &\text{s.t } \ A\mathbf{x} = \mathbf{b}, \quad A \in \mathbb{R}^{p \times n} \end{aligned}$$

נפתור (נבחין כי צריך רק שוויון אז נשתמש רק בכופל לגראנז' ν):

- תחילה נכתוב את הלגראנז'יאן

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\nu}) &= \mathbf{x}^T \mathbf{x} + \sum_{i=1}^{p} \nu_i \cdot (\mathbf{a}_i^T \mathbf{x} - b_i) \\ &= \mathbf{x}^T \mathbf{x} + \boldsymbol{\nu}^T \cdot (A\mathbf{x} - \mathbf{b}) \end{aligned}$$

- כעת נכתוב את $g(\boldsymbol{\nu})$:

$$(*) \quad g(\boldsymbol{\nu}) = \inf_{\mathbf{x}} \mathbf{x}^T \mathbf{x} + \boldsymbol{\nu}^T \cdot (A\mathbf{x} - \mathbf{b})$$

- נמצא אינפימום על ידי גזירה והשוואה לאפס:

$$\begin{aligned} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\nu}) &= 2\mathbf{x} + A^T \boldsymbol{\nu} = \mathbf{o} \\ &\Rightarrow \mathbf{x} = -\frac{1}{2} A^T \boldsymbol{\nu} \end{aligned}$$

- נציב את הפתרון ב - $(*)$ ונקבל את $g(\boldsymbol{\nu})$:

$$\begin{aligned} g(\boldsymbol{\nu}) &= -\frac{1}{4} \boldsymbol{\nu}^T \left(AA^T\right) \boldsymbol{\nu} + \boldsymbol{\nu}^T \left(-\frac{1}{2} AA^T \boldsymbol{\nu} - \mathbf{b}\right) \\ &= -\frac{1}{4} \boldsymbol{\nu}^T \left(AA^T\right) \boldsymbol{\nu} - \boldsymbol{\nu}^T \mathbf{b} \\ &\leq \underbrace{\inf_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{x} \mid A\mathbf{x} = \mathbf{b}\right)}_{\mathbf{p}^*} \end{aligned}$$

3.3 The Lagrange Dual Problem - הבעיה הדואלית

אמרנו שהפונקציה הדואלית מהווה חסם תחתון לערך המינימלי של בעית האופטימיזציה שלנו.
השאלה היא מה הוא הערך **הכי טוב** שניתן לקבל מאותה פונקציה דואלית.
הראינו קודם כי לכל $\boldsymbol{\lambda} \geq 0$ ולכל $\boldsymbol{\nu}$:

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathbf{p}^*$$

אך אנחנו רוצים למצוא חסם שהוא כמה שיותר הדוק לפי $\mathbf{p}^*$.
לפיכך נוכל לפתור את:

$$\begin{aligned} &\text{maximize} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &\text{s.t} \boldsymbol{\lambda} \geq 0 \end{aligned}$$

ל- $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ אשר פותרים את הבעיה הנ"ל נקרא **dual optimal** ול-$\mathbf{x}$ האופטימלי נקרא **primal optimal**

הערה: הבעיה הדואלית **תמיד תהיה קונבקסית**

דוגמה: תכנון לינארי - standard form Linear Programming

$$\begin{aligned} &\text{minimize } \mathbf{c}^T\mathbf{x} \\ &\text{s.t} A\mathbf{x} = \mathbf{b} \\ &\mathbf{x} \geq \mathbf{0} \end{aligned}$$

כאשר נבחין כי האילוץ השני הפוך ממה שאנחנו מכירים - הכפלה באפס תהפוך את האילוץ השני ל:

$$-\mathbf{x} \leq \mathbf{0}$$

- תחילה נרשום את הלגראנז'יאן:

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) &= \mathbf{c}^T\mathbf{x} - \boldsymbol{\lambda}^T\mathbf{x} + \boldsymbol{\nu}^T(A\mathbf{x} - \mathbf{b}) \\ &= -\boldsymbol{\nu}^T\mathbf{b} + (\mathbf{c}^T - \boldsymbol{\lambda}^T + \boldsymbol{\nu}^T A)\mathbf{x} \end{aligned}$$

- כיוון שאין חסם נבחין כי:

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\boldsymbol{\nu}^T\mathbf{b} & , \mathbf{c} - \boldsymbol{\lambda} - A^T\boldsymbol{\nu} = \mathbf{0} \\ -\infty & , \mathbf{c} - \boldsymbol{\lambda} - A^T\boldsymbol{\nu} \neq \mathbf{0} \end{cases}$$

- נוכל לעבור לבעיה יותר פשוטה המוגדרת על פי ν:

$$\begin{aligned} &\underset{\nu,\boldsymbol{\lambda}}{max} \boldsymbol{\nu}^T\mathbf{b} \\ &\text{s.tc} - \boldsymbol{\lambda} + A^T\boldsymbol{\nu} = \mathbf{0} \\ &\boldsymbol{\lambda} \geq 0 \end{aligned}$$

3.4 דואליות חלשה - Weak Duality

כל מה שדיברנו עליו עד כה הוא בעצם **דואליות חלשה** - מצב בו הערך האופטימלי של הבעיה הדואלית שמסומן ב- $\mathbf{d}^*$ הוא חסם תחתון לערך האופטימלי של הבעיה הפרימלית:

$$\mathbf{d}^* \leq \mathbf{p}^*$$

הגדרה: להפרש $\mathbf{p}^* - \mathbf{d}^*$ נהוג לקרוא **duality gap** והוא תמיד אי-שלילי.

3.5 דואליות חזקה - Strong Duality

דואליות חזקה, בניגוד לדואליות חלשה היא מקרה בו פתרון המקסימום שמתקבל בבעיה הדואלית שווה ממש לפתרון המינימום המתקבל מהבעיה הפרימלית:

$$\mathbf{d}^* = \mathbf{p}^*$$

הגדרה: **בעיה קונבקסית** - היא בעית אופטימיזציה בה:

(1) קיימת פונקציית מטרה $f_0(\mathbf{x})$ שהיא פונקציה קונבקסית
(2) ישנם אילוצי אי-שוויון $f_i(\mathbf{x})$ אשר f_i קונבקסית לכל $i = 1, .., m$
(3) יש אילוצי שוויון שהם אפיניים (לינאריים)

משפט סלייטר - Slater's Theorem

אם נתונה בעיה קונבקסית (f_0 קונבקסית, f_i קונבקסית לכל i)

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}) \\ &\text{s.t} \quad f_i(\mathbf{x}) \leq 0 \quad i = 1, .., m \\ &\qquad A\mathbf{x} = \mathbf{b} \end{aligned}$$

וקיים $\mathbf{x}$ עבורו:

$$f_i(\mathbf{x}) < 0, \quad i = 1, .., m \quad A\mathbf{x} = \mathbf{b}$$

אזי:

- קיימת דואליות חזקה
- ה-duality gap הוא 0

הרחבה למשפט:

משפט: בנוסף לתנאי משפט סלייטר, אם קיים k עבורו $f_1, .., f_k$ הן פונקציות אפיניות אז תתקיים דואליות חזקה אם קיים $\mathbf{x}$ כך ש:

- $f_i(\mathbf{x}) \leq 0, \quad i = 1, .., k$
- $f_i(\mathbf{x}) < 0, \quad i = k+1, .., m$
- $A\mathbf{x} = \mathbf{b}$

הערה: בד"כ נתעסק במצבים בהם יש מספר פונקציות אפיניות (ההרחבה למשפט).

3.6 Complementary Slackness

נניח ש- $\mathbf{x}^*$ הוא הפתרון האופטימלי לבעיה הפרימלית ו- $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ הוא הזוג המביא למקסימום את הבעיה הדואלית (dual optimal point) ונניח שמתקיימת דואליות חזקה, אז:

$$\begin{aligned} f_0(\mathbf{x}^*) &\overset{(1)}{=} g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &\overset{(2)}{=} \inf_{\mathbf{x}\in\mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right) \\ &\overset{(3)}{\leq} f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}^*) \\ &\overset{(4)}{\leq} f_0(\mathbf{x}^*) \end{aligned}$$

(1) דואליות חזקה
(2) הגדרת הפונקציה הדואלית
(3) הגדרת האינפימום
(4) הגדרת האילוצים + מההנחה ש $\mathbf{x}^*$ הוא הפתרון האופטימלי

מסקנה: מכך שהתחלנו ב $f_0(\mathbf{x}^*)$ וגם סיימנו בו - נסיק כי לאורך כל הדרך נוכל לשים שוויונים ובפרט בין שלבים 2 ו-3 וגם 3 ו-4. כמו-כן:

(1) מהשוויון בין (2) ל- (3):
 (א) $\mathbf{x}^*$ (האופטימום של הבעיה הפרימלית) הוא גם המינימייזר של הלגראנז'יאן
 (ב) מהשוויון בין (3) ל- (4) ומהאילוץ ש$\sum_{i=1}^p \nu_i h_i(\mathbf{x}^*) = 0$:
  (i) בהכרח $\sum_{i=1}^m \lambda_i f_i(\mathbf{x}^*) = 0$
   (א') בפרט $\lambda_i f_i(\mathbf{x}^*) = 0$ **לכל** $i = 1, .., m$ - (**complementary slackness**)
   (ב') $\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0$ או $f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$

3.7 KKT (Karush-Kuhn-Tucker) optimality conditions

נניח ש- $\mathbf{x}^*$ הוא הפתרון האופטימלי לבעיה הפרימלית ו- $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ הוא הזוג המביא למקסימום את הבעיה הדואלית (dual optimal point) ונניח שמתקיימת דואליות חזקה, אז במידה ו f_i גזירה לכל i נסיק את תנאי KKT:

Karush-Kuhn-Tucker (KKT) conditions

- האילוצים הפרימלים חייבים להתקיים:

$$f_i(\mathbf{x}^*) \leq 0, \;\; i = 1, .., m \quad h_i(\mathbf{x}^*), \;\; i = 1, .., p$$

- האילוצים הדואליים חייבים להתקיים:

$$\lambda_i^* \geq 0, \quad i = 1, .., m$$

- ה- complementary slackness חייב להתקיים:

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \quad i = 1, .., m$$

- הגרדיאנט של הלגראנז'יאן חייב להתאפס בנקודה $\mathbf{x}^*$:

$$\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}^*) = 0$$

3.8 KKT conditions for convex problems

אם ידוע שהבעיה המקורית היא קונבקסית משמע f_i קונבקסי ו h_i אפינית לכל i - מספיק למצוא פתרונות כלשהם:

$$\tilde{\mathbf{x}}, \quad \left(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}\right)$$

אשר מקיימים את תנאי KKT ואז נסיק ש:

- $\mathbf{x}$ - הוא **פרימלי אופטימלי** (אופטימום לבעיה הפרימלית)
- $\left(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}\right)$ - הוא **דואלי** אופטימלי (ממקסמים את הבעיה הדואלית)
- מתקיימת דואליות חזקה

4 הבעיה הדואלית של ה-SVM

ניזכר בבעיה:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &\text{s.t} \quad y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1, \ \ n = 1,..,N \end{aligned}$$

נבחין כי זוהי פונקצית מטרה (f_0) קונבקסית עם אילוצי אי-שוויון אפיניים ולכן ממה שהגדרנו עד כה נסיק שישנה **דואליות חזקה**

- נחשב את הלגראנז'יאן - שוב לא לשכוח להפוך את סימן האילוץ ל- $1 - y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right) \leq 0$:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n \cdot \left(1 - y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right)\right)$$

- נרצה להביא למינימום לפי **$\mathbf{w}$ ולפי** b:
  - לפי b ו-$\mathbf{w}$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = \mathbf{0}$$

$$\iff \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

- נכתוב את g המתאימה עם הצבת $\mathbf{w}$ ו- b שמצאנו:

$$g(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m + \sum_{n=1}^{N} \alpha_n$$

- נרצה לפתור את הבעיה הדואלית של ה- SVM:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^{N} \alpha_n \\ &\text{s.t} \ \ \boldsymbol{\alpha} \geq 0 \\ &\qquad \boldsymbol{\alpha}^T\mathbf{y} = 0 \end{aligned}$$

- כעת נוכל להמיר את האילוצים ל - Quadratic programming ולקבל $\boldsymbol{\alpha}^* \leftarrow QP(Q_0, \mathbf{p}_0, A_0, \mathbf{c})$
  - נמצא ש:

$$\mathbf{w}^* = \sum_{n=1}^{N} \alpha_n^* y_n \mathbf{x}_n$$

- נרצה גם למצוא b על מנת להשלים את הפתרון
  - על פי Complementary slackness נוכל להסיק:

$$\alpha_n \cdot \left(1 - y_n \cdot \left(\mathbf{w}^T \mathbf{x}_n + b\right)\right) = 0$$

  - מכך שהנחנו שאכן קיימות נקודות שצריך להפריד - בהכרח לא יכול להיות ש $\mathbf{0} = \mathbf{w}^*$ ולכן בהכרח קיים $s \in \{1, .., N\}$ כך ש- $\alpha_s > 0$.
  - ובפרט אם קיים s עבורו $\alpha_s > 0$ הרי ש- $y_s \cdot \left(\mathbf{w}^{*T} \mathbf{x}_s + b^*\right) = 1$
    - ⋆ לכן:

$$y_s \cdot \left(\mathbf{w}^{*T} \mathbf{x}_s + b^*\right) = 1$$

$$\overset{\backslash y_s}{\iff} \mathbf{w}^{*T} \mathbf{x}_s + b^* = y_s$$

$$\iff b^* = y_s - \mathbf{w}^{*T} \mathbf{x}_s$$

הערה: **חשוב מאוד** - הנקודות שנמצאות על שולי ה-margin נקראות **support vectors** ובהכרח כל נקודה x_n עבורה $\alpha_n > 0$ היא בהכרח support vector

ניתן להסיק שהחזאי שלנו **מוגדר** אך ורק על ידי support vectors כיוון שרק הם מוגדרים!!!

בפרט ניתן להגדיר את החזאי בתור:

$$g(\mathbf{x}) = sign(\mathbf{w^*}^T \mathbf{x} + b) = sign\left(\sum_{\alpha_s > 0} \alpha_s^* y_s \mathbf{x}_s^T \mathbf{x} + b^*\right)$$