תרגיל בית 4 – מבוא ללמידה חישובית

מגיש: נתן בלוך

.Theory Questions

שאלה 1. SGD With Projection.

אלגוריתם $gradient\ descent\ with\ projection$ כולל –
$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla f(\boldsymbol{x}_t)$$
$$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{K}}(\boldsymbol{y}_{t+1})$$
כאשר $\Pi_{\mathcal{K}}(y) := arg \min_{x\in\mathcal{K}} ||\boldsymbol{x}-\boldsymbol{y}||$.

סעיף א. נרצה לפתור את הבעיה הפרימאלית של SVM הנתונה ע"י –
$$\frac{1}{2}||w||^2 + \frac{C}{m}\sum_{i=1}^{m} \max\{0\,, 1 - y_i \boldsymbol{w}\cdot\boldsymbol{x_i}\}$$
תחת האילוץ של נורמה חסומה, כלומר כאשר $\mathcal{K} = \{x : ||x|| \le R\}$.

פיתרון. ראשית נבחין כי החישוב של $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{K}}(\boldsymbol{y}_{t+1})$ מהווה בעצם בעיית אופטימיזציה, המוגדרת ע"י –
$$\min_{x\in\mathcal{K}} ||\boldsymbol{x}-\boldsymbol{y}_{t+1}|| \qquad which\ is\ equivalent\ to \qquad \min_{x} ||\boldsymbol{x}-\boldsymbol{y}_{t+1}|| \quad \boldsymbol{s.t.} \quad ||\boldsymbol{x}|| \le R$$
נבצע אבחנה ולפיה אם $||\boldsymbol{y}_{t+1}|| \le R$, אזי $\boldsymbol{y}_{t+1}$ **הינו הפיתרון לבעיית אופטימזציה זו** שכן הוא משיג את המינימום של פונק' המטרה, שהינו כאן באפס, שכן מתקיים בעצם כי $||\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}|| = ||0|| = 0$, וכן $x = \boldsymbol{y}_{t+1} \in \mathcal{K}$, כלומר האילוצים מתקיימים ופונק' המטרה משיגה את המינימום שלה.

לכן, כיוון שנרצה לחשב את ה-$projection$, נפתור את בעיית האופטימיזציה הבאה – $\min_{x} ||\boldsymbol{x}-\boldsymbol{y}_{t+1}|| \quad \boldsymbol{s.t.} \quad ||\boldsymbol{x}|| \le R$, ונבחין שכיוון שמתקיים $R \ge 0$, $\left||\boldsymbol{x}|\right|$, אזי $||\boldsymbol{x}|| \le R \iff ||\boldsymbol{x}||^2 \le R^2$, ולכן נפתור את בעיית האופטמיזציה הבאה, שהינה שקולה –
$$\min_{x} ||\boldsymbol{x}-\boldsymbol{y}|| \qquad \boldsymbol{s.t.} \quad ||\boldsymbol{x}||^2 \le R^2$$
הלגרנג'יאן של הבעיה הינו – $\mathcal{L}(x,a) = \| \boldsymbol{x}-\boldsymbol{y} \|^2 + a(\| \boldsymbol{x} \|^2 - R^2)$. הגרדיאנט לפי x –
$$\nabla_{\mathrm{x}}\, \mathcal{L}(x,a) = (2a+2)x - 2\boldsymbol{y}$$
וע"י השוואת הגרדיאנט לאפס, נקבל –
$$\nabla_{\mathrm{x}}\, \mathcal{L}(\boldsymbol{x},a) = 0 \implies (2a+2)\boldsymbol{x} - 2\boldsymbol{y} = 0 \implies (a+1)\boldsymbol{x} = \boldsymbol{y} \implies \boldsymbol{x} = \frac{1}{a+1}\boldsymbol{y}$$
הבעיה הדואלית מוגדרת ע"י $\max_{a} g(a)$, כאשר $g(a) = \min_{x} \mathcal{L}(x,a)$, ולכן כיוון ש-$\mathcal{L}(x,a)$ הינה קמורה(בסכום של פונק' קמורות ולינאריות), אזי נקודת הקיצון של הפונק'(בה הגרדיאנט מתאפס) תהיה נקודת המינימום המוחלט של הפונק'. על כן, כעת נציב זאת ב-$\mathcal{L}(x,a)$ כדי לקבל –
$$\mathcal{L}(x,a) = \| \boldsymbol{x}-\boldsymbol{y} \|^2 + a(\| \boldsymbol{x} \|^2 - R^2) = \| \frac{1}{a+1}\boldsymbol{y} - \boldsymbol{y} \|^2 + a\left(\| \frac{1}{a+1}\boldsymbol{y} \|^2 - R^2\right) =$$
$$= \left(\frac{a}{a+1}\right)^2 \| \boldsymbol{y} \|^2 + a\left(\left(\frac{1}{a+1}\right)^2 \| \boldsymbol{y} \|^2 - R^2\right) = \frac{a^2}{(a+1)^2} \| \boldsymbol{y} \|^2 + a\left(\frac{1}{(a+1)^2} \| \boldsymbol{y} \|^2 - R^2\right) =$$
$$= \frac{a^2+a}{(a+1)^2} \| \boldsymbol{y} \|^2 - aR^2 = \frac{a(a+1)}{(a+1)^2} \| \boldsymbol{y} \|^2 - aR^2 = \frac{a}{a+1} \| \boldsymbol{y} \|^2 - aR^2$$
ולכן הבעיה הדואלית הינה הבעיה הבאה – $\boxed{\begin{matrix} \max_{a} \ \frac{a}{a+1} \| \boldsymbol{y} \|^2 - aR^2 \\ s.t. \quad a \ge 0 \end{matrix}}$.

וכעת נפתור אותה ע"י גזירה לפי a והשוואת הנגזרת לאפס(שכן מדובר בפונק' במשתנה יחיד, כאשר $\| \boldsymbol{y} \|^2, R^2$ הינם קבועים). נקבל –
$$g^{'}(\alpha) = \| \boldsymbol{y} \|^2 \left(\frac{1\cdot(a+1) - 1\cdot a}{(a+1)^2}\right) - R^2 = \| \boldsymbol{y} \|^2 \left(\frac{-1}{(a+1)^2}\right) - R^2$$
וע"י השוואת $g'(\alpha) = 0$, נקבל $\| \boldsymbol{y} \|^2 \left(\frac{1}{(a+1)^2}\right) - R^2 = 0 \implies \frac{1}{(a+1)^2} = \frac{R^2}{\|\boldsymbol{y}\|^2} \implies (a+1)^2 = \frac{\|\boldsymbol{y}\|^2}{R^2} \implies \boxed{\boldsymbol{a} = \frac{\|\boldsymbol{y}\|}{\boldsymbol{R}} - \boldsymbol{1}}$

וניתן לוודא כי זו אכן נקודת מקסימום ע"י גזירה פעם נוספת של הפונקציה, ואכן נקבל כי $g^{''}\left(\frac{\|\boldsymbol{y}\|}{R} - 1\right) < 0$ –. כלומר, הראנו כי המקסימום של $g(a)$ מתקבל בנקודה $a = \frac{\|\boldsymbol{y}\|}{R} - 1$. כמו כן, הראנו עד כה כי לפי תנאי KKT, נקבל – $\boldsymbol{x}^* = \frac{1}{a^*+1}\boldsymbol{y}$, וכיוון שהראנו כי מהתוכנית הדואלית מתקבל הביטוי הבא עבור הפיתרון האופטימלי $a^* = \frac{\|\boldsymbol{y}\|}{R} - 1$, נקבל את השיוויון הבא – $\boldsymbol{x}^* = \frac{1}{a^*+1}\boldsymbol{y} = \frac{1}{\frac{\|\boldsymbol{y}\|}{R} - 1 + 1}\boldsymbol{y} = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\cdot R$, ובסה"כ $\boxed{\boldsymbol{x}^* = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\cdot \boldsymbol{R}}$.

כלומר, הראנו כי לכל y, מתקיים –

$$\boxed{\boldsymbol{x}^* = \frac{\boldsymbol{y}}{\| \boldsymbol{y} \|} \cdot \boldsymbol{R} = arg \min_{x \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}||}$$

תחת האילוץ של נורמה חסומה, כלומר כאשר $\mathcal{K} = \{x : ||x|| \leq R\}$.

על כן, ניתן לשנות את אלגוריתם ה-SGD כך שיכיל את ה-$projection\ step$, באופן הבא –

$$\boldsymbol{x}_{t+1} = \begin{cases} \boldsymbol{y}_{t+1} & if \quad \| \boldsymbol{y}_{t+1} \| \leq R \\ \frac{\boldsymbol{y}_{t+1}}{\| \boldsymbol{y}_{t+1} \|} \cdot R & else \end{cases}$$

כנדרש בסעיף א.

סעיף ב. נתונה קבוצה קמורה $\mathcal{K}$, נקודה $y \in \mathbb{R}^d$, ונסמן $x = \Pi_{\mathcal{K}}(y)$. צריך להראות שלכל $z \in \mathcal{K}$, מתקיים –

$$\| \mathrm{y} - \mathrm{z} \| \geq \| \mathrm{x} - \mathrm{z} \|$$

פיתרון. יהי $z \in \mathcal{K}$ כלשהו. נחלק את ההוכחה למספר מקרים אפשריים.
מקרה 1. אם $y \in \mathcal{K}$, אזי נוכל להסיק כי $\boldsymbol{x} = \Pi_{\mathcal{K}}(\boldsymbol{y}) = \boldsymbol{y}$, שכן $\boldsymbol{0} = \| \boldsymbol{0} \| = \| \boldsymbol{y} - \boldsymbol{y} \| \underbrace{\leq}_{\substack{as \\ y \in \mathcal{K}}} \min_{x \in \mathcal{K}} \| \boldsymbol{x} - \boldsymbol{y} \|$, ולכן מאי-שליליות הנורמה, נקבל כי בהכרח $\min_{x \in \mathcal{K}} \| \boldsymbol{x} - \boldsymbol{y} \| = 0$, ולכן נוכל להסיק כי $\boldsymbol{y} = arg \min_{x \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}||$. כלומר, מהגדרה, נסיק כי $\boldsymbol{x} = \Pi_{\mathcal{K}}(\boldsymbol{y}) = \boldsymbol{y}$. במקרה זה, בוודאי כי מתקיים השיוויון – $\| \mathrm{x} - \mathrm{z} \| = \| \mathrm{y} - \mathrm{z} \|$, ובפרט מתקיים גם אי-שיוויון $\| \mathrm{y} - \mathrm{z} \| \geq \| \mathrm{x} - \mathrm{z} \|$ כפי שנדרש.

מקרה 2. אם $y \notin \mathcal{K}$, אזי נוכל להסיק כי $x \neq y$, שכן $y \notin \mathcal{K}$ וכן $x \in \mathcal{K}$. **אם $\boldsymbol{z} = \boldsymbol{x}$,** אזי נוכל להסיק כי בהכרח מתקיים –

$$\| \mathrm{y} - \mathrm{z} \| \geq 0 = \| 0 \| = \| \mathrm{x} - \mathrm{x} \| \underbrace{=}_{\substack{as \\ x=z}} \| \mathrm{x} - \mathrm{z} \|$$

כנדרש. **לכן נותר להוכיח את נכונות הטענה במקרה בו $\boldsymbol{x} \neq \boldsymbol{z}$.** נבחין כי לכן מדובר ב-3 נקודות שונות. נסתכל על המשולש המוגדר ע"י הנקודות x, y, z. נוכיח כי הצלע המוגדרת ע"י הנקודות y, z הינה הגדולה ביותר במשולש, ולכן בהכרח נקבל את אי-השיוויון הנדרש - $\| \mathrm{y} - \mathrm{z} \| \geq \| \mathrm{x} - \mathrm{z} \|$.

- נניח בשלילה כי הצלע בין x ל-y היא הצלע הגדולה ביותר במשולש, ולכן $\| \mathrm{y} - \mathrm{x} \| \geq \| \mathrm{y} - \mathrm{z} \|$. נניח בשלילה כי $\| \mathrm{y} - \mathrm{z} \| = \| \mathrm{y} - \mathrm{x} \|$, כלומר זהו משולש שווה שוקיים תחת הנחות אלו. נעביר אנך מהנקודה y לישר העובר בין z ל-x, ונסמן את נקודת החיתוך ב-a. משיקולים גיאומטריים, הנקודה a נמצאת **בין** הנקודות x ו-z, ולכן מקמירות $\mathcal{K}$ קיים $t \in (0,1)$ כך ש-$a = (1-t)z + tx$, וכן מקמירות $a \in \mathcal{K}$.
  כמו כן, נסתכל על המשולש המוגדר ע"י הנקודות y, x, a, ונבחין כי זהו משולש ישר-זוית שכן הוגדר ע"י האנך לישר. על כן, נוכל להסיק כי הצלע הגדולה ביותר במשולש הינה הצלע המוגדרת ע"י y, x, כלומר מתקיים – $\| \mathrm{y} - a \| < \| \mathrm{y} - \mathrm{x} \|$, וזה בסתירה לכך ש-
  $$x = \Pi_{\mathcal{K}}(y) := arg \min_{x \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}||$$
  שכן מצאנו כי הנקודה $a \in \mathcal{K}$, מקיימת $\| \mathrm{y} - a \| < \| \mathrm{y} - \mathrm{x} \|$, **וזו סתירה** לבחירת x כנקודה הקרובה ביותר אל y הנמצאת ב-$\mathcal{K}$.
  לפיכך, נסיק כי לא ייתכן שהצלע הגדולה ביותר במשולש המוגדר ע"י הנקודות x, y, z, הינה הצלע בין בין x ל-y.
- נניח בשלילה שהצלע בין x ל-z היא הצלע הגדולה במשולש, ולכן $\| \mathrm{x} - \mathrm{z} \| \geq \| \mathrm{y} - \mathrm{z} \|$, נניח בשלילה כי $\| \mathrm{y} - \mathrm{z} \| = \| \mathrm{x} - \mathrm{z} \|$, כלומר זהו משולש שווה שוקיים תחת הנחות אלו. נעביר אנך מהנקודה y לישר העובר בין z ל-x, ונסמן את נקודת החיתוך ב-a. משיקולים גיאומטריים, הנקודה a נמצאת **בין** הנקודות x ו-z, ולכן מקמירות $\mathcal{K}$ קיים $t \in (0,1)$ כך ש-$a = (1-t)z + tx$, וכן מקמירות $a \in \mathcal{K}$.
  כמו כן, נסתכל על המשולש המוגדר ע"י הנקודות y, x, a, ונבחין כי זהו משולש ישר-זוית שכן הוגדר ע"י האנך לישר. על כן, נוכל להסיק כי הצלע הגדולה ביותר במשולש הינה הצלע המוגדרת ע"י y, x, כלומר מתקיים – $\| \mathrm{y} - a \| < \| \mathrm{y} - \mathrm{x} \|$, וזו סתירה לבחירת x כך שיתקיים $x = \Pi_{\mathcal{K}}(y)$, כלומר סתירה לכך ש-x הנקודה הקרובה ביותר ל-y הנמצאת ב-$\mathcal{K}$, שכן קיבלנו כי $a \in \mathcal{K}$ והינה קרובה יותר ל-y, מאשר x.
  על כן, הגענו לסתירה לבחירת x כנקודה הקרובה ביותר אל y הנמצאת ב-$\mathcal{K}$. לפיכך, נסיק כי לא ייתכן שהצלע הגדולה ביותר במשולש המוגדר ע"י הנקודות x, y, z, הינה הצלע בין x ל-z.

מכאן, נסיק כי בהכרח הצלע הגדולה ביותר במשולש המוגדר ע"י הנקודות x, y, z, הינה הצלע המוגדרת ע"י הנקודות y, z הינה הגדולה ביותר במשולש, ולכן בהכרח נקבל את אי-השיוויון הנדרש - $\| \mathrm{y} - \mathrm{z} \| \geq \| \mathrm{x} - \mathrm{z} \|$, כרצוי. ■

סעיף ג. נוכיח כי משפט ההתכנסות של $\boldsymbol{Gradient\ Descent}$ מתקיים.

תהי $f: \mathbb{R}^d \to \mathbb{R}$ פונק' דיפרנציאבלית וקמורה. נסמן גם $x^* = argmin_x\, f(x)$, ונניח גם שמתקיים – $\|x^*\| \le B$. בנוסף, נניח כי הגרדיאנט חסום ע"י G, כלומר לכל $x \in \mathbb{R}^d$, מתקיים כי – $\|\nabla f(x)\| \le G$. נסמן גם $\bar{x} = \frac{1}{T}\sum_{i=1}^{T} x_i$.

יהי $\varepsilon > 0$ כלשהו. נראה כי עבור צעדים קבועים $\eta_t = \eta = \frac{\epsilon}{G^2}$, ועבור $T = \frac{B^2G^2}{\varepsilon^2}$, נקבל כי – $f(\bar{x}) - f(x^*) \le \varepsilon$.

פיתרון. נגדיר את נקודת ההתחלה כ-$x_1 = 0$, נקודת הראשית של $\mathbb{R}^d$.

נתון כי f קמורה, ולכן מתקיים עבורה – $f(\bar{x}) = f\left(\frac{1}{T}\sum_{i=1}^{T} x_i\right) \underbrace{\le}_{jensen.} \frac{1}{T}\sum_{i=1}^{T} f(x_i)$. כמו כן, מתקיים –

$$f(\bar{x}) - f(x^*) \le \frac{1}{T}\sum_{i=1}^{T} f(x_i) - \frac{T}{T} f(x^*) = \frac{1}{T}\sum_{i=1}^{T} f(x_i) - \frac{1}{T}\sum_{i=1}^{T} f(x^*) = \frac{1}{T}\sum_{i=1}^{T}\big(f(x_i) - f(x^*)\big) \underbrace{\le}_{convexity.} \frac{1}{T}\sum_{i=1}^{T} \nabla f(x_i)\cdot(x_i - x^*)$$

כעת, נסתכל על הביטוי $\|x_{i+1} - x^*\|_2^2$ כפונק' פוטנציאל לגבי התכנסות $\{x_i\}$ אל הנקודה הנדרשת x^*. מתקיים לפי הגדרת $SGD\ with\ Projection$,

$$\|x_{i+1} - x^*\|_2^2 = \|\Pi_{\mathcal{K}}(y_{i+1}) - x^*\|_2^2 \underbrace{\le}_{\substack{by\ (b)\\ for\ z=x^*}} \|y_{i+1} - x^*\|_2^2 = \|x_i - \eta\nabla f(x_i) - x^*\|_2^2 =$$

$$= \|x_i - x^*\|_2^2 - 2\eta\nabla f(x_i)(x_i - x^*) + \eta^2\|\nabla f(x_i)\|_2^2$$

וע"י העברת אגפים וחלוקה בגורם של 2η נקבל –

$$(*)\quad \nabla f(x_i)(x_i - x^*) = \frac{1}{2\eta}\|x_i - x^*\|_2^2 - \frac{1}{2\eta}\|x_{i+1} - x^*\|_2^2 + \frac{\eta}{2}\|\nabla f(x_i)\|_2^2$$

וכעת, ע"י הצבה בחסם שקיבלנו ולפיו –

$$f(\bar{x}) - f(x^*) \le \frac{1}{T}\sum_{i=1}^{T}\nabla f(x_i)\cdot(x_i - x^*) \underbrace{=}_{(*)} \frac{1}{T}\sum_{i=1}^{T}\Big(\frac{1}{2\eta}\|x_i - x^*\|_2^2 - \frac{1}{2\eta}\|x_{i+1} - x^*\|_2^2 + \frac{\eta}{2}\|\nabla f(x_i)\|_2^2\Big) =$$

$$= \underbrace{\frac{1}{2\eta T}\sum_{i=1}^{T}\big(\|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2\big)}_{telescopic\ series.} + \frac{\eta}{2T}\sum_{i=1}^{T}\underbrace{\|\nabla f(x_i)\|_2^2}_{\le G} \le \frac{1}{2\eta T}\Big\|\underbrace{x_1 - x^*}_{=-x^*}\Big\|_2^2 - \underbrace{\frac{1}{2\eta T}\|x_{T+1} - x^*\|_2^2}_{\ge 0} + \underbrace{\frac{\eta}{2T}G^2}_{\substack{<\frac{\eta}{2}G^2\\ as\ T\ge 1}} \le$$

$$\le \frac{1}{2\eta T}\underbrace{\|x^*\|_2^2}_{\le B^2} + \frac{\eta}{2}\underbrace{G^2}_{=\frac{\varepsilon}{\eta}} \le \frac{1}{2\eta T}\underbrace{B^2}_{=\frac{\varepsilon^2 T}{G^2}} + \underbrace{\frac{\eta}{2}\cdot\frac{\varepsilon}{\eta}}_{=\frac{\varepsilon}{2}} = \frac{1}{2\eta T}\cdot\frac{\varepsilon^2 T}{G^2} + \frac{\varepsilon}{2} \underbrace{=}_{G^2=\frac{\varepsilon}{\eta}} \frac{1}{2\eta}\cdot\frac{\varepsilon^2}{\frac{\varepsilon}{\eta}} + \frac{\varepsilon}{2} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

כנדרש, שכן הראנו כי – $\boldsymbol{f(\bar{x}) - f(x^*) \le \varepsilon}$, ולכן אכן ניתן להסיק כי משפט ההתכנסות של $\boldsymbol{Gradient\ Descent}$ מתקיים.

שאלה 2. SVM With Multiple Classes.

סעיף א. נסמן ב- $w_1^*(\beta), w_2^*(\beta)$ את הפיתרון לבעיית ה-$multiclass$ עם $K = 2$, כאשר גם $penalty\ is\ \beta$. נסמן גם $w^*(\beta')$ את הפיתרון לבעיית ה-SVM הסטנדרטית עם $penalty\ of\ \beta'$.

פיתרון. ראשית נראה כי $w_1^*(\beta) = -w_2^*(\beta)$. נניח בשלילה כי $w_1^*(\beta) \neq -w_2^*(\beta)$, ונגדיר – $v_1 := \frac{w_1^*(\beta)-w_2^*(\beta)}{2}$ וכן $v_2 := -v_1 = \frac{w_2^*(\beta)-w_1^*(\beta)}{2}$.
נבחין כי $v_1 - v_2 = w_1^*(\beta) - w_2^*(\beta)$, ולכן נוכל להסיק(לפי העובדה הנתונה) כי מתקיים – $\ell(v_1, v_2, x_i, y_i) = \ell(w_1^*(\beta), w_2^*(\beta), x_i, y_i)$,
לכל $i \in \{1, \dots, n\}$, ונסמן ב-(*) נתונים אלה. כעת, מתקיים –

$$\|v_1\|^2 + \|v_2\|^2 = \|v_1\|^2 + \|-v_1\|^2 = 2\cdot\|v_1\|^2 = 2\cdot\left\|\frac{w_1^*(\beta)-w_2^*(\beta)}{2}\right\|^2 = \frac{1}{2}\cdot\|w_1^*(\beta)-w_2^*(\beta)\|^2 \qquad (\mathbf{1})$$

כמו כן, מתקיים –

$$\frac{1}{2}\cdot\|w_1^*(\beta)-w_2^*(\beta)\|^2 < \|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2 \iff$$

$$\iff \frac{1}{2}\cdot(w_1^*(\beta)-w_2^*(\beta))\cdot(w_1^*(\beta)-w_2^*(\beta)) < \|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2 \iff$$

$$\iff \frac{1}{2}\|w_1^*(\beta)\|^2 - w_1^*(\beta)\cdot w_2^*(\beta) + \frac{1}{2}\|w_2^*(\beta)\|^2 < \|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2 \iff$$

$$\iff 0 < \frac{1}{2}\|w_1^*(\beta)\|^2 + w_1^*(\beta)\cdot w_2^*(\beta) + \frac{1}{2}\|w_2^*(\beta)\|^2 \iff$$

$$\iff 0 < \frac{1}{2}\|w_1^*(\beta)\|^2 + w_1^*(\beta)\cdot w_2^*(\beta) + \frac{1}{2}\|w_2^*(\beta)\|^2 \iff$$

$$\iff 0 < \|w_1^*(\beta)\|^2 + 2w_1^*(\beta)\cdot w_2^*(\beta) + \|w_2^*(\beta)\|^2 \iff$$

$$\iff 0 < \left(w_1^*(\beta) + w_2^*(\beta)\right)^2 \iff \boldsymbol{w_1^*(\beta) \neq -w_2^*(\beta)}$$

ולפי הנחת השלילה לפיה $w_1^*(\beta) \neq -w_2^*(\beta)$, נקבל כי אכן – $\frac{1}{2}\cdot\|w_1^*(\beta)-w_2^*(\beta)\|^2 < \|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2$, ובנוסף כיוון שהראנו לפי השיוויון $(\mathbf{1})$, כי מתקיים – $\|v_1\|^2 + \|v_2\|^2 = \frac{1}{2}\cdot\|w_1^*(\beta)-w_2^*(\beta)\|^2$, **נוכל להסיק לכן כי –** $\|\boldsymbol{v_1}\|^2 + \|\boldsymbol{v_2}\|^2 < \|\boldsymbol{w_1^*(\beta)}\|^2 + \|\boldsymbol{w_2^*(\beta)}\|^2$.
כאשר נסמן א"ש זה ב-$(\mathbf{2})$. בעת, נקבל כי מתקיים –

$$f(v_1, v_2) = \frac{\beta}{2}(\|v_1\|^2 + \|v_2\|^2) + \frac{1}{n}\sum_{i=1}^{n}\ell(v_1, v_2, x_i, y_i) \underbrace{<}_{\substack{by\\(2)}} \frac{\beta}{2}(\|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2) + \frac{1}{n}\sum_{i=1}^{n}\ell(v_1, v_2, x_i, y_i) =$$

$$\underbrace{=}_{\substack{by\ the\\fact\ (*)}} \frac{\beta}{2}(\|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2) + \frac{1}{n}\sum_{i=1}^{n}\ell(w_1^*(\beta), w_2^*(\beta), x_i, y_i) = f\left(w_1^*(\beta), w_2^*(\beta)\right)$$

כלומר, הראנו כי מתקיים –$f(v_1, v_2) < f\left(w_1^*(\beta), w_2^*(\beta)\right)$, בסתירה לכך ש-$w_1^*(\beta), w_2^*(\beta)$ פיתרון אופטימלי, ובפרט משיגים את המינימום של f.

על כן, נסיק כי הנחת השלילה ולפיה $w_1^*(\beta) \neq -w_2^*(\beta)$ הינה שגויה, ולכן בהכרח $w_1^*(\beta) = -w_2^*(\beta)$.

כעת, נסמן ב- $w_1^*(\beta), w_2^*(\beta), \beta$ פיתרון אופטימלי של בעיית ה-$multiclass\ SVM$. לפי מה שהוכחנו עד כה, נסיק כי $w_1^*(\beta) = -w_2^*(\beta)$.
כעת, נגדיר את הפיתרון הבא לבעיית ה-$binary\ classification\ SVM$, ונראה כי הוא אופטימלי. נגדיר – $w^*(\beta') = 2w_2^*(\beta)$ וכן נגדיר $\beta' = \frac{\beta}{2}$.
נראה כי עבור הגדרה זו של הפרמטרים נקבל פיתרון השקול לפיתרון האופטימלי של בעיית ה-$multiclass\ SVM$, ולפיכך נסיק זהו פיתרון אופטימלי.
כמו כן, נעיר כי עבור בעיית ה-$multiclass\ SVM$, מתקיים כי $y_i \in \{1,2\}$, ואילו עבור בעיית ה-$binary\ classification\ SVM$, מדובר בפרמטרים מסווגים של $\hat{y}_i \in \{-1,1\}$. כמו כן, נבחין כי מתקיים הקשר הבאה ביניהם – $\hat{y}_i = (-1)^{y_i}$. ובמילים, מתקיים כי $y_i = 1$ מתאים ל-$\hat{y}_i = -1$, וכן גם $y_i = 2$ מתאים ל-$\hat{y}_i = 1$.

נעשה ראשית מספר אבחנות.

אבחנה 1. מכך שהראנו שמתקיים – $w_1^*(\beta) = -w_2^*(\beta)$, ניתן להסיק כי $\|w_1^*(\beta)\| = \|w_2^*(\beta)\|$, ולפיכך גם – $\|w_1^*(\beta)\|^2 = \|w_2^*(\beta)\|^2$ $(*)$.

אבחנה 2. נחשב את פונ' ההפסד תחת ההנחה $w_1^*(\beta) = -w_2^*(\beta)$. נחלק לשני מקרים לפי ערכו של $y_i \in \{1,2\}$.

אם $y_i = 1$, אזי –

$$\ell(w_1^*(\beta), w_2^*(\beta), x_i, y_i) = \ell(w_1^*(\beta), -w_1^*(\beta), x_i, 1) = \max\left\{\underbrace{w_1^* \cdot x_i - w_1^* \cdot x_i + \mathbb{I}[1 \neq 1]}_{=0}, -w_1^* \cdot x_i - w_1^* \cdot x_i + \mathbb{I}[2 \neq 1]\right\} =$$

$$= \max\{0, -2w_1^* \cdot x_i + 1\} = \max\{0, 1 - 2w_1^* \cdot x_i\} \underbrace{=}_{\substack{as \\ y_i=1}} \max\{0, 1 + (-1)^{y_i} 2w_1^* \cdot x_i\}$$

אם $y_i = 2$, אזי –

$$\ell(w_1^*(\beta), w_2^*(\beta), x_i, y_i) = \ell(w_1^*(\beta), -w_1^*(\beta), x_i, 2) = \max\left\{w_1^* \cdot x_i - (-w_1^*) \cdot x_i + \mathbb{I}[1 \neq 2], \underbrace{-w_1^* \cdot x_i - (-w_1^*) \cdot x_i + \mathbb{I}[2 \neq 2]}_{=0}\right\} =$$

$$= \max\{0, 2w_1^* \cdot x_i + 1\} = \max\{0, 1 + 2w_1^* \cdot x_i\} \underbrace{=}_{\substack{as \\ y_i=2}} \max\{0, 1 + (-1)^{y_i} 2w_1^* \cdot x_i\}$$

ובבל כללי, ניתן לכתוב –

$$\boxed{\boldsymbol{\ell(w_1^*(\beta), w_2^*(\beta), x_i, y_i) = \max\{0, 1 + (-1)^{y_i} 2w_1^* \cdot x_i\}}}$$

ונסמן שיוויון זה ב-$(**)$.

נסתכל כעת על פונק' המטרה של בעיית ה-$binary\ classification\ SVM$, עבור $w^*(\beta')$ ו- β' –

$$\frac{\beta'}{2}\|w^*(\beta')\|^2 + \frac{1}{n}\sum_{i=1}^{n}\max\{0, 1 - \hat{y}_i w^*(\beta') \cdot x_i\} = \frac{\beta}{4}\|2w_2^*(\beta)\|^2 + \frac{1}{n}\sum_{i=1}^{n}\max\{0, 1 - \hat{y}_i 2w_2^*(\beta) \cdot x_i\} =$$

$$= \beta\|w_2^*(\beta)\|^2 + \frac{1}{n}\sum_{i=1}^{n}\max\{0, 1 + (-1)^{y_i} 2w_1^*(\beta) \cdot x_i\} \underbrace{=}_{(*)} \frac{\beta}{2}(\|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2) + \frac{1}{n}\sum_{i=1}^{n}\max\{0, 1 + (-1)^{y_i} 2w_1^*(\beta) \cdot x_i\} =$$

$$\underbrace{=}_{(**)} \frac{\beta}{2}(\|w_1^*(\beta)\|^2 + \|w_2^*(\beta)\|^2) + \frac{1}{n}\sum_{i=1}^{n}\ell(w_1^*(\beta), w_2^*(\beta), x_i, y_i)$$

וזו כאמור פונק' המטרה של בעיית ה-$multiclass\ SVM$, וכיוון שבחרנו את $w_1^*(\beta), w_2^*(\beta), \beta$ כפיתרון האופטימלי של בעיית ה-$multiclass\ SVM$, נוכל להסיק למעשה כי הפיתרון של $w^*(\beta') = 2w_2^*(\beta)$ וכן $\beta' = \frac{\beta}{2}$, הינו הפיתרון האופטימלי של בעיית ה-$binary\ classification\ SVM$, כנדרש.

ובסה"כ, לכן, אם הפיתרון האופטימלי לבעיית ה-$\boldsymbol{multiclass\ SVM}$, מסומן ע"י $\boldsymbol{w_1^*(\beta), w_2^*(\beta), \beta}$, אזי פיתרון המוגדר ע"י –

$$\boxed{\boldsymbol{w^*(\beta') = 2w_2^*(\beta), \quad \beta' = \frac{\beta}{2}}}$$

הינו הפיתרון האופטימלי של בעיית ה-$\boldsymbol{binary\ classification\ SVM}$.

סעיף ב. לא יודע.

סעיף ג. נניח כי $\beta = 0$, ולכן נניח כי המקרה הינו ה-*seperable case*, כלומר קיימים $w_1^*, \ldots, w_K^*$ כך שמתקיים עבורם – $y_i = argmax_y w_y^* \cdot x_i$.
נדרש להראות שלכל *minimizer* של $f(w_1, \ldots, w_K)$, נקבל שגיאה של אפס על המסווג.

פיתרון. יהיו $w_1, \ldots, w_K$ כך שמדובר ב-*minimizer* של f. במקרה בו $\beta = 0$, נקבל כי –

$$f(w_1, \ldots, w_K) = \underbrace{\frac{\beta}{2}\sum_{j=1}^{K}\|w_j\|^2}_{=0} + \frac{1}{n}\sum_{i=1}^{n}\ell(w_1, \ldots, w_K, x_i, y_i) = \frac{1}{n}\sum_{i=1}^{n}\ell(w_1, \ldots, w_K, x_i, y_i)$$

נרצה להראות כי $\ell(w_1, \ldots, w_K, x_i, y_i) = 0$ לכל $1 \le i \le n$. לשם כך מספיק להראות כי – $f(w_1, \ldots, w_K) = \frac{1}{n}\sum_{i=1}^{n}\ell(w_1, \ldots, w_K, x_i, y_i) = 0$, ולכן **לשם כך מספיק להראות כי – $\sum_{i=1}^{n}\ell(w_1, \ldots, w_K, x_i, y_i) = 0$.** מספיק להראות זאת כיוון ש-ℓ אי-שלילית, ולכן אם הסכום הנ"ל הינו 0, אזי גם כך כל גורם בסכום יהיה שווה לאפס.
אם כך, לפי ההנחה, מתקיים – $y_i = argmax_y\, w_y^* \cdot x_i$, ולכן ניתן להסיק כי – $w_{y_i}^* \cdot x_i \ge w_j^* \cdot x_i$ לכל $j \in [K]$. מכאן, נגדיר לכל $i \in [K]$ את –

$$\varepsilon_i = \max_{\substack{j \in [K] \\ s.t.\ j \ne y_i}} w_j^* \cdot x_i - w_{y_i}^* \cdot x_i$$

ולכן לפי ההגדרה, ולפי ההנחות, נסיק כי – $\varepsilon_i \le 0$, לכל $i \in [K]$. נגדיר גם $\boxed{\boldsymbol{\varepsilon = \min_i |\varepsilon_i|}}$. כעת, נגדיר את $w_1', \ldots, w_K'$ באופן הבא – $\boxed{\boldsymbol{w_i' = \frac{1}{\varepsilon} w_i^*}}$.

כעת, הנחנו כי $w_1, \ldots, w_K$ הינם *minimizer* של f, ולכן מתקיים –

$$f(w_1, \ldots, w_K) \le f(w_1', \ldots, w_K') \quad \Rightarrow \quad \underbrace{\frac{\beta}{2}\sum_{j=1}^{K}\|w_j\|^2}_{\substack{=0 \\ as\ \beta=0}} + \frac{1}{n}\sum_{i=1}^{n}\ell(w_1, \ldots, w_K, x_i, y_i) \le \underbrace{\frac{\beta}{2}\sum_{j=1}^{K}\|w_j'\|^2}_{\substack{=0 \\ as\ \beta=0}} + \frac{1}{n}\sum_{i=1}^{n}\ell(w_1', \ldots, w_K', x_i, y_i)$$

$$\Rightarrow \sum_{i=1}^{n}\ell(w_1, \ldots, w_K, x_i, y_i) \le \sum_{i=1}^{n}\ell(w_1', \ldots, w_K', x_i, y_i) = \sum_{i=1}^{n}\max_{j \in [K]}\{w_j' \cdot x_i - w_{y_i}' \cdot x_i + \mathbb{I}[j \ne y_i]\} =$$

$$[for\ j = y_i \quad in\ the\ \max objective\ we\ get\ value\ 0] = \sum_{i=1}^{n}\max\left\{0\,, \max_{\substack{j \in [K] \\ j \ne y_i}}\{w_j' \cdot x_i - w_{y_i}' \cdot x_i + \underbrace{\mathbb{I}[j \ne y_i]}_{=1}\}\right\} =$$

$$\left[w_i' = \frac{1}{\varepsilon}w_i^*\right] = \sum_{i=1}^{n}\max\left\{0\,, 1 + \max_{\substack{j \in [K] \\ j \ne y_i}}\{\frac{1}{\varepsilon}w_j^* \cdot x_i - \frac{1}{\varepsilon}w_{y_i}^* \cdot x_i\}\right\} = \sum_{i=1}^{n}\max\left\{0\,, 1 + \frac{1}{\varepsilon}\overbrace{\max_{\substack{j \in [K] \\ j \ne y_i}}\{w_j^* \cdot x_i - w_{y_i}^* \cdot x_i\}}^{=\varepsilon_i}\right\} =$$

$$= \sum_{i=1}^{n}\max\left\{0\,, 1 + \frac{1}{\varepsilon}\varepsilon_i\right\} = \quad [\boldsymbol{\varepsilon_i \le 0} \quad \Rightarrow \boldsymbol{\varepsilon_i = -|\varepsilon_i|}] \quad = \sum_{i=1}^{n}\max\left\{0\,, 1 + \frac{-|\varepsilon_i|}{\min_i |\varepsilon_i|}\right\} = \sum_{i=1}^{n}\max\left\{0\,, 1 - \underbrace{\frac{|\varepsilon_i|}{\min_i |\varepsilon_i|}}_{\le -1}\right\} =$$

$$\le \sum_{i=1}^{n}\max\{0\,, 1 - 1\} = \sum_{i=1}^{n} 0 = 0$$

כך שקיבלנו –

$$\sum_{i=1}^{n}\ell(w_1', \ldots, w_K', x_i, y_i) \le 0$$

ומאי-שליליות נוכל להסיק כי – $\ell(w_1, \ldots, w_K, x_i, y_i) = 0$ לכל $1 \le i \le n$, כנדרש על מנת להוכיח כי כל *minimizer* של f משיג שגיאה של אפס.

שאלה 3. Kernel Implementation of Primal SGD.

פיתרון. עדכון ה-SGD המתאים הינו –

$$w_{t+1} = (1-\eta)w_t + b_t \eta C y_i \phi(x_i)$$

כאשר מניחים כי $b_t = 1$. נקבל באינדוקציה(ההוכחה טכנית בלבד ונובעת מהצבת השיוויונים זה בזה) את השיוויון הבא –

$$w_{t+1} = (1-\eta)^{t+1} w_0 + \sum_{j=0}^{t} (1-\eta)^j \eta C y_i^j \phi\left(x_i^j\right)$$

כאשר כאן נסמן את הדגימה שנבחרה בשלב ה- j באופן הבא – $\left(x_i^j, y_i^j\right)$. כמו כן, ע"י בחירה של $w_0 = 0$, נקבל כי את הביטוי הבא –

$$w_{t+1} = \sum_{j=0}^{t} (1-\eta)^j \eta C y_i^j \phi\left(x_i^j\right)$$

כעת, נאתחל את המערך A להיות מערך של אפסים. התא $A[i]$ יכיל את המקדמים של הדגימה (x_i, y_i) בביטוי w_{t+1}. נבחין כי ניתן לחשב את ערכי המערך במהלך ריצת האלגוריתם בזמן קבוע ע"י הוספת הערך $(1-\eta)^j \eta C y_i^j$ באיטרציה ה-j לתא המייצג את הדגימה $\left(x_i^j, y_i^j\right)$. כעת, נקבל את השיוויון

$$w_{t+1} = \sum_{i=1}^{n} A[i] \phi(x_i)$$

כאשר n הינו מספר הדגימות ב-$data\ points$, ולפיכך בסה"כ נדרש לשמור כאן $O(n)$ מספרים שמייצגים את המקדמים של $\phi(x_i)$ בביטוי w_{t+1}.
כמו כן, במעבר מ-w_t אל w_{t+1} נוסף מקדם נוסף עבור $\phi\left(x_i^{t+1}\right)$ וניתן לכן להוסיף לערך בתא ה-$A[i]$ את הביטוי $(1-\eta)^{t+1} \eta C y_i^{t+1}$ שנוסף במעבר מ-w_t אל w_{t+1}. על כן, כל עדכון בזמן קבוע(כלומר $O(1)$, ובפרט בזמן $O(n)$ כנדרש).

כלומר, נבחין כי לא ניתן לייצג את w_t כוקטור שכן ייתכן ומדובר בוקטור ממרחב במימד אינסופי, ועל כן נייצג את w_t ע"י n פרמטרים כפי שהוסבר.
כעת, נראה כיצד ניתן לבצע סיווג של נקודה כלשהי. נסמן ב-w את תוצאת ה-SGD שהתקבלה, וכן נסמן ב-A את מערך המקדמים המייצג את w, בדומה להסבר קודם לכן. בהנתן נקודה x, נרצה לסווג אותה, אך נדרש להשתמש בערך ה-$feature$, כלומר ב-$\phi(x)$.
הסיווג מתבצע ע"י חישוב באופן הבא –

$$prediction = sign(\langle w, \phi(x)\rangle) = sign\left(\left(\sum_{i=1}^{n} A[i]\phi(x_i)\right)\phi(x)\right) = sign\left(\sum_{i=1}^{n} A[i] \underbrace{\phi(x_i)\phi(x)}_{\substack{Kernel\\ trick.}}\right) = sign\left(\sum_{i=1}^{n} A[i] K(x_i, x)\right)$$

סה"כ הראנו כי נדרש $O(n)$ זיכרון על מנת לייצג את w_t ולבצע SGD, וכן הראנו שכל עדכון מתבצע בזמן קבוע(ובפרט יותר מהר מ-$O(n)$), כנדרש בשאלה. כמו כן, הראנו כיצד ניתן לבצע סיווג של נקודה בעזרת הקרנל טריק.

כנדרש.

שאלה 4. Weighted SVM.

נתונים משקלים $0 \le v_i \le 1$, עבור $i \in \{1,..,n\}$. בעיית האופטימיזציית הינה –

$$\boxed{\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n} v_i\,\xi_i \\ \boldsymbol{s.t.}\quad & y_i(w\cdot x_i\ +\ b) \ge 1-\xi_i \quad \forall i \in \{1,\dots,n\} \\ & \xi_i \ge 0 \qquad \forall i \in \{1,\dots,n\}\end{aligned}}$$

סעיף א. נמצא את הלגרנג'יאן.

פיתרון. הבעיה הנתונה שקולה לבעיה (הסטנדרטית) הבאה –

$$\boxed{\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n} v_i\,\xi_i \\ \boldsymbol{s.t.}\quad & 1-\xi_i - \ y_i(w\cdot x_i\ +\ b) \le 0 \quad \forall i \in \{1,\dots,n\} \\ & -\xi_i \le 0 \qquad \forall i \in \{1,\dots,n\}\end{aligned}}$$

לכן, הלגרנג'יאן הינו –

$$\boxed{\mathcal{L}(w,b,\xi,a,r) = \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n} v_i\xi_i + \sum_{i=1}^{n} a_i\big(1-\xi_i - \ y_i(w\cdot x_i\ +\ b)\big) - \sum_{i=1}^{n} r_i\xi_i}$$

סעיף ב. נמצא את התוכנית הדואלית.

פיתרון. נחקור ראשית $\min_{w,b,\xi} \ \mathcal{L}(w,b,\xi,a,r)$ ע"י השוואת הגרדיאנט של $\mathcal{L}(w,b,\xi,a,r)$ לפי w,b,ξ לאפס.

עבור w_k, נקבל –

$$\frac{\partial\mathcal{L}(w,b,\xi,a,r)}{\partial w_k} = w_k - \sum_{i=1}^{n} a_i y_i x_i^{(k)}$$

וע"י $\frac{\partial\mathcal{L}(w,b,\xi,a,r)}{\partial w_k} = 0$, נקבל כי – $w_k = \sum_{i=1}^{n} a_i y_i x_i^{(k)}$. כמו כן, נבחין שמתקיים –

$$\boxed{\mathrm{w} = \sum_{i=1}^{n} a_i y_i x_i}$$

עבור b, נקבל –

$$\frac{\partial\mathcal{L}(w,b,\xi,a,r)}{\partial b} = -\sum_{i=1}^{n} a_i y_i$$

וע"י $\frac{\partial\mathcal{L}(w,b,\xi,a,r)}{\partial b} = 0$, נקבל כי – $\boxed{\sum_{i=1}^{n} a_i y_i = 0}$. כמו כן, נבחין שאם $\sum_{i=1}^{n} a_i y_i \neq 0$, לא קיים מינימום ללגרנג'יאן, שכן במקרה זה ניתן לקחת את הלגרנג'יאן ל-$-\infty$. לכן נדרוש בתוכנית הדואלית אילוץ לפיו $\sum_{i=1}^{n} a_i y_i = 0$.

עבור ξ_k, נקבל –

$$\frac{\partial\mathcal{L}(w,b,\xi,a,r)}{\partial \xi_k} = Cv_k - a_k - r_k$$

וע"י $\frac{\partial\mathcal{L}(w,b,\xi,a,r)}{\partial \xi_k} = 0$, נקבל כי – $\boxed{a_k = Cv_k - r_k}$, וכן $\boxed{r_k = Cv_k - a_k}$, וכן $\boxed{Cv_k = r_k + a_k}$.

נבחין כי בתוכנית הדואליות נקבל אילוצי אי-שליליות לפיהם $a_i, r_i \ge 0$, אך כאן מצאנו כי $a_i = Cv_i - r_i$, ולכן ניתן להחליף את האילוץ של $r_i \ge 0$ ע"י התנאי של $\boxed{\boldsymbol{a_i \le Cv_i}}$. הנכונות של טיעון זה נובעת מכך שאם $a_i \le Cv_i$, אזי $0 \le Cv_i - a_i = r_i$.

הלגרנג'יאן $\mathcal{L}(w,b,\xi,a,r)$ הינה פונק' קמורה, כסכום של פונק' קמורות, והמינימום שלה לפי w,b,ξ מתקבל בנקודות בהן הגרדיאנט מתאפס. כמו כן, הפונקצייה הדואלית מוגדרת באופן הבא – $g(a,r) = \min_{w,b,\xi} \ \mathcal{L}(w,b,\xi,a,r)$, וכן פונק' המטרה של התוכנית הדואלית הינה - $\max_{a,r} \ g(a,r)$.
על כן, נרצה למצוא את $g(a,r)$ ולכן נציב את האילוצים הללו בלגרנג'יאן.

כעת, נסתכל על $\mathcal{L}(w,b,\xi,a,r)$ ונציב את האילוצים הללו –

$$\mathcal{L}(w,b,\xi,a,r)=\frac{1}{2}||w||_2^2+C\sum_{i=1}^{n}v_i\xi_i+\sum_{i=1}^{n}a_i\big(1-\xi_i-\ y_i(w\cdot x_i\ +\ b)\big)-\sum_{i=1}^{n}r_i\xi_i=$$

$$=\frac{1}{2}||w||_2^2+C\sum_{i=1}^{n}v_i\xi_i+\sum_{i=1}^{n}a_i-\sum_{i=1}^{n}a_i\xi_i-w\underbrace{\sum_{i=1}^{n}a_i\,y_ix_i}_{=w}-b\underbrace{\sum_{i=1}^{n}a_iy_i}_{=0}-\sum_{i=1}^{n}r_i\xi_i=$$

$$=-\frac{1}{2}w\cdot w+\sum_{i=1}^{n}a_i+\sum_{i=1}^{n}Cv_i\xi_i-\sum_{i=1}^{n}a_i\xi_i-\sum_{i=1}^{n}r_i\xi_i\underbrace{=}_{Cv_i=(r_i+a_i)}$$

$$=-\frac{1}{2}w\cdot w+\sum_{i=1}^{n}a_i+\sum_{i=1}^{n}(r_i+a_i)\xi_i-\sum_{i=1}^{n}a_i\xi_i-\sum_{i=1}^{n}r_i\xi_i=$$

$$=-\frac{1}{2}\left(\sum_{i=1}^{n}a_iy_ix_i\right)\left(\sum_{i=1}^{n}a_iy_ix_i\right)+\sum_{i=1}^{n}a_i+\underbrace{\sum_{i=1}^{n}r_i\xi_i+\sum_{i=1}^{n}a_i\xi_i-\sum_{i=1}^{n}a_i\xi_i-\sum_{i=1}^{n}r_i\xi_i}_{=0}=$$

$$=\boxed{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_iy_ia_jy_jx_i\cdot x_j+\sum_{i=1}^{n}a_i}$$

ובסה"כ קיבלנו כי התוכנית הדואלית הינה –

$$\boxed{\begin{aligned}&\max_{a,r}\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_iy_ia_jy_jx_i\cdot x_j+\sum_{i=1}^{n}a_i\\&s.t.\ \sum_{i=1}^{n}a_iy_i=0\\&\qquad Cv_i\ge a_i\ge 0\qquad \forall i\in\{1,\dots,n\}\end{aligned}}$$

כאשר נבחין כי r אינו מופיע בפונק' המטרה ובאילוצים(כלומר אינו מופיע בבעית האופטימיזציה), ולכן ניתן להסתכל על בעיית האופטימיזציה לפי a בלבד.

כאשר בעצם $\max\limits_{a,r} g(a)=\min\limits_{a,r}-g(a)$, ולכן התוכנית הדואלית הבאה הינה שקולה גם כן –

$$\boxed{\begin{aligned}&\min_{a,r}\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_iy_ia_jy_jx_i\cdot x_j-\sum_{i=1}^{n}a_i\\&s.t.\ \sum_{i=1}^{n}a_iy_i=0\\&\qquad Cv_i\ge a_i\ge 0\qquad \forall i\in\{1,\dots,n\}\end{aligned}}$$

סעיף ג. נניח בעת כי פתרנו את התוכנית הדואלית. נמצא את $\mathbf{w}^*,\mathbf{b}^*,\boldsymbol{\xi}^*$ שהם פיתרון אופטימלי לתוכנית הפרימאלית.

נסמן את הפיתרון של התוכנית הדואלית באופן הבא – $\boldsymbol{a}^*,\boldsymbol{r}^*$. ניתן למצוא את $\mathbf{w}^*$ באופן מיידי על ידי שימוש בכך שמתקיים –

$$\boxed{\mathbf{w}^*=\sum_{i=1}^{n}\boldsymbol{a_i^*y_ix_i}}$$

על מנת למצוא את $\boldsymbol{\xi}^*$, נשתמש בתכונת $complementary\ slackness$ ולפיה –

$$\forall i\in\{1,\dots,n\}.\quad a_i^*\big(1-\xi_i^*-\ y_i(w^*\cdot x_i\ +\ b^*)\big)=0\quad and\quad r_i^*\xi_i^*=0$$

נמצא את $\boldsymbol{\xi_i}^*$, באופן הבא –

- אם $a_i^*=0$, אזי $r_i^*=Cv_i-a_i^*=Cv_i>0$, ולכן לפי $r_i^*\xi_i^*=0$, נקבל כי בהכרח $\xi_i^*=0$.
- אם $Cv_i>a_i^*>0$, אזי $r_i^*=Cv_i-a_i^*>0$, ולכן לפי $r_i^*\xi_i^*=0$, נקבל כי בהכרח $\xi_i^*=0$.
  במקרה זה, גם ניתן לחשב את b^* ע"י שימוש ב- $a_i^*\big(1-\xi_i^*-\ y_i(w^*\cdot x_i\ +\ b^*)\big)=0$, כך שנקבל –
  $b^*=y_i-w^*\cdot x_i$
- אם $a_i^*=Cv_i\neq 0$, כאשר את המקרים הללו נחשב לאחר מציאת b^*(בהנתן b^*), וכאן לפי $a_i^*\big(1-\xi_i^*-\ y_i(w^*\cdot x_i\ +\ b^*)\big)=0$, נקבל –
  $\xi_i^*=1-\ y_i(w^*\cdot x_i\ +\ b^*)$

שאלה 5. All Points Are Support Vectors.

נתונה קבוצה $S = \{x_i, y_i\}_{i=1}^{2n}$ כך ש- $x_i \in \mathbb{R}^2$, וכן $y_i \in \{-1,1\}$. כמו כן, נתון –

$$x_i = \begin{cases} (1, i) & i \le n \\ (-1, i-n) & i > n \end{cases} \qquad\qquad y_i = \begin{cases} 1 & i \le n \\ -1 & i > n \end{cases}$$

נפתור את בעיית ה-SVM הזו באופן ישיר ונראה כי כל נקודה הינה $support\ vector$.

פיתרון. ראינו בהרצאה כי בעיית האופטימיזציה של SVM הינה –

$$\boxed{\begin{array}{c} \min_{w,b} \frac{1}{2}||w|| \\ \boldsymbol{s.t.}\quad y_i(w \cdot x_i + b) \ge 1 \quad \forall i \in [1,2n] \end{array}}$$

ונסמן גם – $w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$, ונציב את הנתונים כדי לקבל את בעיית האופטימיזצייה הבאה –

$$\boxed{\begin{array}{ll} & \min_{w,b} \frac{1}{2}\sqrt{w_1^2 + w_2^2} \\ \boldsymbol{s.t.} & 1 \cdot \left((w_1, w_2) \cdot \begin{pmatrix} 1 \\ i \end{pmatrix} + b \right) \ge 1 \quad \forall i \in [1, n] \\ & (-1) \cdot \left((w_1, w_2) \cdot \begin{pmatrix} -1 \\ i-n \end{pmatrix} + b \right) \ge 1 \quad \forall i \in [n+1, 2n] \end{array}}$$

ובעיית האופטימיזצייה זו שקולה לבעיית האופטימיזצייה הבאה –

$$\boxed{\begin{array}{ll} & \min_{w,b} \frac{1}{2}\sqrt{w_1^2 + w_2^2} \\ \boldsymbol{s.t.} & w_1 + i \cdot w_2 + b \ge 1 \quad \forall i \in [1, n] \\ & w_1 + (n-i) \cdot w_2 - b \ge 1 \quad \forall i \in [n+1, 2n] \end{array}}$$

ונבחין כי אם $i = n + k$ עבור $1 \le k \le n$, מתקיים - $n - i = n - (n+k) = -k$. לכן בעיית האופטימיזציה שקולה לבעיית האופטימיזצייה הבאה –

$$\boxed{\begin{array}{ll} & \min_{w,b} \ \frac{1}{2}\sqrt{w_1^2 + w_2^2} \\ \boldsymbol{s.t.} & w_1 + i \cdot w_2 + b \ge 1 \quad \forall i \in [1, n] \\ & w_1 - i \cdot w_2 - b \ge 1 \quad \forall i \in [1, n] \end{array}}$$

נבחין כי **עבור כל פיתרון פיזיבילי $\boldsymbol{w, b}$ לבעיית אופטימיזציה זו**, מתקיים עבור w, b - (ע"י חיבור אי-השיוויונות הנ"ל)

$$w_1 + i \cdot w_2 + b + w_1 - i \cdot w_2 - b = 2w_1 \ge 1 + 1 = 2 \quad \Longrightarrow \quad 2w_1 \ge 2 \quad \Longrightarrow \quad \boldsymbol{w_1 \ge 1}$$

הראנו כי כל פיתרון(פיזיבילי) לבעיה זו, מקיים $w_1 \ge 1$, ובפרט הפיתרון האופטימלי לכן. לכן, לכל פיתרון פיזיבילי לבעיה, פונק' המטרה בהכרח מקיימת –

$$\frac{1}{2}\sqrt{\underbrace{w_1^2}_{\ge 1} + \underbrace{w_2^2}_{\ge 0}} \ge \frac{1}{2}\sqrt{1+0} = \frac{1}{2}$$

כלומר, פונק' המטרה חסומה מלמטה ע"י $\frac{1}{2}$. כעת, נסתכל על הפיתרון $\boxed{\boldsymbol{w_1 = 1, w_2 = 0, b = 0}}$, ונראה כי הוא מקיים את האילוצים ומשיג את החסם התחתון של פונק' המטרה, ולכן הוא בהכרח פיתרון אופטימלי לבעיית האופטימיזציה הזו. ואכן, האילוצים מתקיימים שכן –

$$\forall i \in [1, n]. \quad w_1 + i \cdot w_2 + b = 1 + i \cdot 0 + 0 \ge 1$$
$$\forall i \in [1, n]. \quad w_1 - i \cdot w_2 - b = 1 - i \cdot 0 - 0 \ge 1$$

כמו כן , פונק' המטרה במקרה זה מקבלת את הערך – $\frac{1}{2}\sqrt{w_1^2 + w_2^2} = \frac{1}{2}\sqrt{1+0} = \frac{1}{2}$, כלומר אכן פונק' המטרה אכן משיגה את החסם התחתון שלה, וכפי שהראנו הפיתרון הנ"ל מקיים את האילוצים(כלומר, פיזיבילי), **ולכן הפיתרון $\boxed{\boldsymbol{w_1 = 1, w_2 = 0, b = 0}}$ הינו בהכרח הפיתרון האופטימלי.**

בנוסף, נדרש להראות כי כל נקודה x_i, y_i הינה $support\ vector$. ואכן, מתקיים –

$$\forall i \in [1, n]. \quad y_i(w \cdot x_i + b) = 1 \cdot \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ i \end{pmatrix} + 0 \right) = \mathbf{1}$$

$$\forall i \in [n+1, n]. \quad y_i(w \cdot x_i + b) = (-1) \cdot \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ i-n \end{pmatrix} + 0 \right) = -1 \cdot -1 = \mathbf{1}$$

ואכן קיבלנו כי כל נקודה $(\boldsymbol{x_i, y_i})$ הינה $\boldsymbol{support\ vector}$, כנדרש.

שאלה 6. Separability Using RBF Kernels.

נבחן את $RBF\ Kernel$ עם פרמטר $\sigma = 1$, וכן קלטים שהם סקלרים, כלומר $x \in \mathbb{R}$. נניח גם כי נתון סט אימון $\{(x_i, y_i)\}_{i=1}^{n}$ כאשר כל ה-x_i שונים.

סעיף א. נדרש להראות כי $hard\ SVM$ עם $RBF\ Kernel$ משיג תמיד שגיאה אפס.

נתונות העובדות הבאות –

(i). מטריצה המוגדרת ע"י $A_{ij} = e^{a_i \beta_j}$ כאשר $a_1 \neq a_2 \neq \cdots \neq a_n$ וכן $\beta_1 \neq \beta_2 \neq \cdots \neq \beta_n$ הינה מדרגה מלאה.
(ii). אם מטריצת הקרנל המתאימה לנתוני סט האימון הינה מדרגה מלאה, אזי $hard\ SVM$ משיג שגיאה של אפס.

פיתרון. קרנל RBF מוגדר באופן הבא - $K(x, x') = e^{-\frac{1}{\sigma^2}(x_i - x_j)^2}$ עבור המקרה של קלטים ב-$\mathbb{R}$. נסמן ב-A את מטריצת הקרנל המתאימה, המוגדרת באופן שראינו בהרצאה – $A_{i,j} = K(x_i, x_j)$. בכתיב מטריצות, ניתן לכתוב ע"י שימוש בהנחה כי הפרמטר הינו $\sigma = 1$ –

$$A = \begin{pmatrix} e^{-(x_1-x_1)^2} & \cdots & e^{-(x_1-x_n)^2} \\ \vdots & \ddots & \vdots \\ e^{-(x_n-x_1)^2} & \cdots & e^{-(x_n-x_n)^2} \end{pmatrix} = \begin{pmatrix} e^{-x_1^2+2x_1x_1-x_1^2} & \cdots & e^{-x_1^2+2x_1x_n-x_n^2} \\ \vdots & \ddots & \vdots \\ e^{-x_n^2+2x_nx_1-x_1^2} & \cdots & e^{-x_n^2+2x_nx_n-x_n^2} \end{pmatrix}$$

נרצה להראות כי מטריצת הקרנל הינה מדרגה מלאה(כלומר דרגה n, כלומר $full\ rank$), ולהשתמש בעובדה (ii) הנתונה בשאלה כדי להסיק שעבור מטריצת קרנל עם דרגה מלאה, $hard\ SVM$ משיג שגיאת אפס.

נסתכל על המטריצה הבאה – $B_{i,j} = e^{a_i \beta_i}$, כאשר $\boldsymbol{a_i = 2x_i}$ וכן $\boldsymbol{\beta_i = x_i}$. נבחין כי $a_1 \neq a_2 \neq \cdots \neq a_n$ שכן נתון כי $x_1 \neq x_2 \neq \cdots \neq x_n$, וכן ניתן להסיק כי גם $\beta_1 \neq \beta_2 \neq \cdots \neq \beta_n$ משיקולים דומים. **לכן, לפי הנתון בעובדה (i), ניתן להסיק כי המטריצה B הינה מדרגה מלאה($full\ rank$).**

בכתיב מטריציוני, נקבל – $B = \begin{pmatrix} e^{2x_1x_1} & \cdots & e^{2x_1x_n} \\ \vdots & \ddots & \vdots \\ e^{2x_nx_1} & \cdots & e^{2x_nx_n} \end{pmatrix}$. **משיקולי אלגברה לינארית, כפל של שורה או עמודה במטריצה בקבוע(שונה מאפס) אינה משנה את הדרגה של המטריצה.** לכן, כעת נכפול כל שורה i במטריצה B בגורם $e^{-x_i^2} \neq 0$, ונקבל את המטריצה הבאה B' –

$$B' = \begin{pmatrix} \boldsymbol{e^{-x_1^2}} e^{2x_1x_1} & \cdots & \boldsymbol{e^{-x_1^2}} e^{2x_1x_n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{e^{-x_n^2}} e^{2x_nx_1} & \cdots & \boldsymbol{e^{-x_n^2}} e^{2x_nx_n} \end{pmatrix} = \begin{pmatrix} e^{-x_1^2+2x_1x_1} & \cdots & e^{-x_1^2+2x_1x_n} \\ \vdots & \ddots & \vdots \\ e^{-x_n^2+2x_nx_1} & \cdots & e^{-x_n^2+2x_nx_n} \end{pmatrix}$$

וביוון שהמטריצה B' התקבלה ע"י כפל בקבועים(שונים מאפס) של כל שורה של המטריצה B, נקבל כי המטריצה B' הינה מדרגה מלאה($full\ rank$).

כעת נכפול כל עמודה i במטריצה B' בגורם $e^{-x_i^2} \neq 0$, ונקבל את המטריצה הבאה B'' –

$$B'' = \begin{pmatrix} \boldsymbol{e^{-x_1^2}} e^{-x_1^2+2x_1x_1} & \cdots & \boldsymbol{e^{-x_n^2}} e^{-x_1^2+2x_1x_n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{e^{-x_1^2}} e^{-x_n^2+2x_nx_1} & \cdots & \boldsymbol{e^{-x_n^2}} e^{-x_n^2+2x_nx_n} \end{pmatrix} = \begin{pmatrix} e^{-x_1^2+2x_1x_1-x_1^2} & \cdots & e^{-x_1^2+2x_1x_n-x_n^2} \\ \vdots & \ddots & \vdots \\ e^{-x_n^2+2x_nx_1-x_1^2} & \cdots & e^{-x_n^2+2x_nx_n-x_n^2} \end{pmatrix} = \begin{pmatrix} e^{-(x_1-x_1)^2} & \cdots & e^{-(x_1-x_n)^2} \\ \vdots & \ddots & \vdots \\ e^{-(x_n-x_1)^2} & \cdots & e^{-(x_n-x_n)^2} \end{pmatrix} = A$$

משיקולים דומים, המטריצה B'' הינה מדרגה מלאה, שכן התקבלה ע"י כפל של עמודות מטריצה מדרגה מלאה בקבועים(שונים מאפס), ופעולות אלה משמרות את דרגת המטריצה. כמו כן, מתקיים כי $B'' = A$ למעשה, ולכן **ניתן להסיק כי מטריצת הקרנל A הינה מדרגה מלאה, כרצוי.**

לפיכך, ע"י שימוש בעובדה (ii), נותר להסיק את הנדרש – $hard\ SVM$ משיג שגיאה של אפס, כנדרש.

סעיף ב. נראה כי התוצאה לא מתקיימת כאשר $x_1 = x_2$.

פיתרון.

נסתכל על מטריצת הקרנל כפי שהגדרנו בסעיף הקודם תחת הפרמטר $\sigma = 1$ –

$$A = \begin{pmatrix} e^{-(x_1-x_1)^2} & \cdots & e^{-(x_1-x_n)^2} \\ \boldsymbol{e^{-(x_2-x_1)^2}} & \cdots & \boldsymbol{e^{-(x_2-x_n)^2}} \\ \vdots & \ddots & \vdots \\ e^{-(x_n-x_1)^2} & \cdots & e^{-(x_n-x_n)^2} \end{pmatrix} \underbrace{\Longrightarrow}_{x_1=x_2} A = \begin{pmatrix} e^{-(x_1-x_1)^2} & \cdots & e^{-(x_1-x_n)^2} \\ \boldsymbol{e^{-(x_1-x_1)^2}} & \cdots & \boldsymbol{e^{-(x_1-x_n)^2}} \\ \vdots & \ddots & \vdots \\ e^{-(x_n-x_1)^2} & \cdots & e^{-(x_n-x_n)^2} \end{pmatrix} \Longrightarrow row\ 1\ is\ equal\ to\ row\ 2$$

$$\Longrightarrow A\ is\ singular \quad \Longrightarrow \quad rank(A) < n$$

לפי הערות המתרגל מהפורום, **ניתן להתייחס לעובדה (ii) כטענת אם ורק אם,** ולכן כאן מצאנו כי מטריצת הקרנל A אינה מדרגה מלאה, ולכן, ניתן להסיק כי התוצאות מהסעיף הקודמות אינן מתקיימות, כלומר במקרה זה, $hard\ SVM$ לא בהכרח ישיג שגיאה של אפס, כנדרש בסעיף זה. ∎

Programming Assignment.

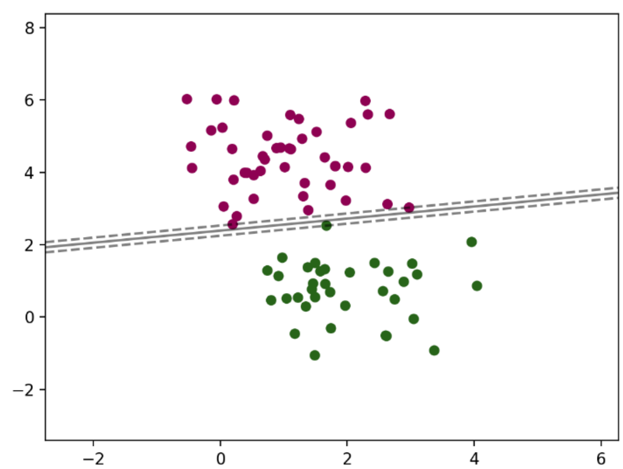
שאלה 1. SVM.

סעיף א. מימוש הפונק' $train\_three\_kernels$ המאמנת 3 מודלים של $Kernel\ SVM$ – ליניארי, ריבועי ו-RBF.
כמו כן, נגדיר פרמטר $penalty\ C = 1000$. **הקוד –**

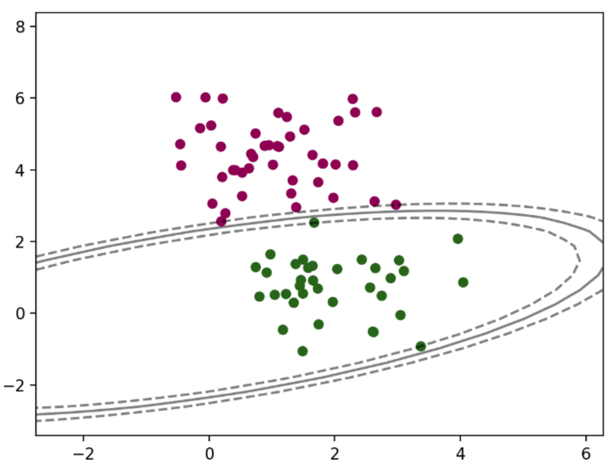
```python
def train_three_kernels(X_train, y_train, X_val, y_val):
    """
    Returns: np.ndarray of shape (3,2):
        A two dimensional array of size 3 that contains the number of support vectors for each class(2) in the three kernels.
    """
    linear_clf = svm.SVC(kernel='linear', C=1000)
    linear_clf.fit(X_train, y_train)

    quadratic_clf = svm.SVC(kernel='poly', C=1000, degree=2)
    quadratic_clf.fit(X_train, y_train)

    rbf_clf = svm.SVC(C=1000)
    rbf_clf.fit(X_train, y_train)

    create_plot(X_train, y_train, rbf_clf)
    plt.show()

    create_plot(X_train, y_train, quadratic_clf)
    plt.show()

    create_plot(X_train, y_train, linear_clf)
    plt.show()

    return np.array([linear_clf.n_support_, quadratic_clf.n_support_, rbf_clf.n_support_])
```

המודלים המתקבלים –

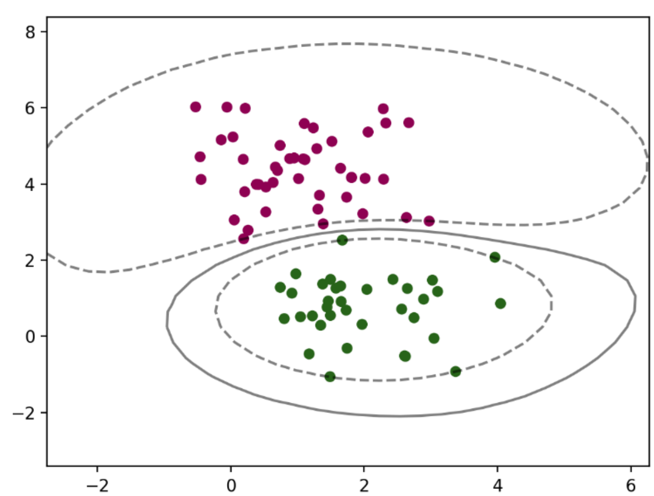
המודל הלינארי



המודל הריבועי



מודל ה-RBF



ניתן לראות כי המודלים שונים זה מזה בסוג המפריד שמתקבל.
המודל הלינארי מפריד עפ"י פונק' לינארית, בעוד המודל הריבועי מפריד עפ"י פונק' ריבועית, וכן מודל ה-RBF הינו הגמיש ביותר וניתן לקבל ע"י פונק' אקספרסיביות($expressive$) יותר.
מספר ה-$Support\ Vectors$ הינו –

$linear: 3\ SVs\ ,\quad quadratic: 4\ SVs,\quad RBF: 6\ SVs$

ואכן הפונק' שכתבנו מחזירה את מספר ה-SVs ומתקבל –

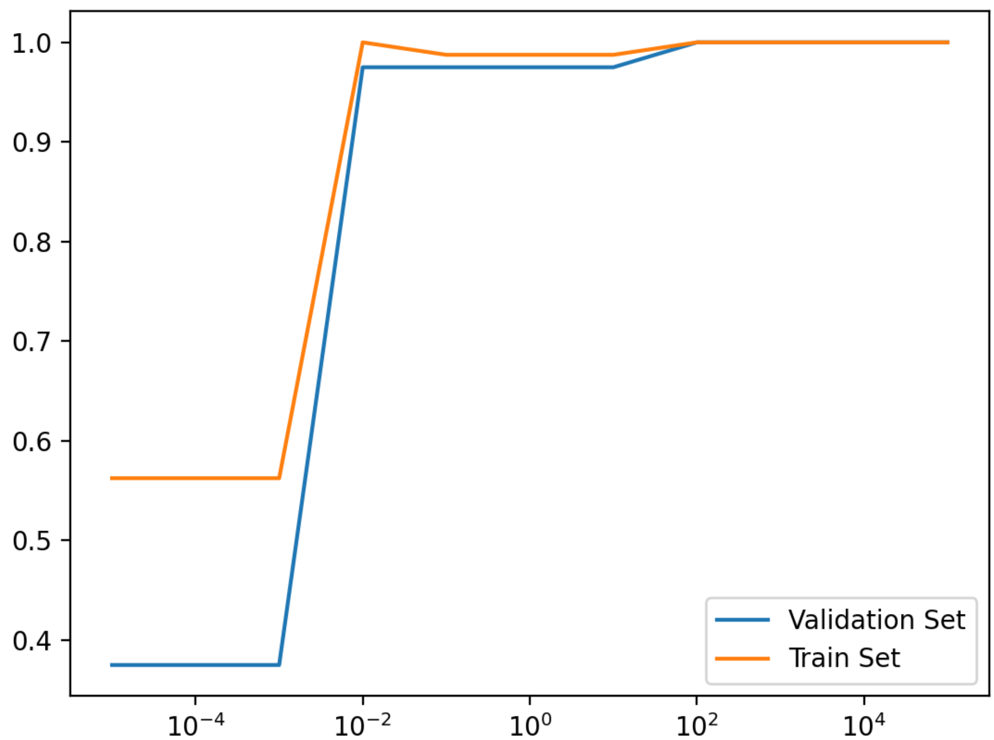
```
# of SVs in linear model: 3
# of SVs in quaratic model: 4
# of SVs in RBF model: 6
```

סעיף ב. מימוש הפונק' $linear\_accuracy\_per\_C$. בסעיף זה נבדוק מספר אפשרויות לערך הפרמטר C ע"י אימון המודל עם ערכי $C = c_i$ שונים, ולאחר מכן בדיקת המודל על סט הואלידציה. נבחר ב-c_i המביא את אחוז הדיוק המירבי על סט ה-$validation$.

הקוד –

```python
def linear_accuracy_per_C(X_train, y_train, X_val, y_val):
    """
        Returns: np.ndarray of shape (11,) :
                    An array that contains the accuracy of the resulting model on the VALIDATION set.
    """
    valid_accuracy = np.zeros(11)
    train_accuracy = np.zeros(11)
    C = [pow(10, i) for i in range(-5, 6)]
    for i in range(11):
        clf = svm.SVC(kernel='linear', C = C[i])
        clf.fit(X_train, y_train)
        valid_accuracy[i] = calc_accuracy(X_val, y_val, clf)
        train_accuracy[i] = calc_accuracy(X_train, y_train, clf)

    valid_line, = plt.plot(C, valid_accuracy)
    train_line, = plt.plot(C, train_accuracy)
    plt.xscale('log')
    plt.legend((valid_line, train_line), ('Validation Set', 'Train Set'))
    plt.show()

    return valid_accuracy
```

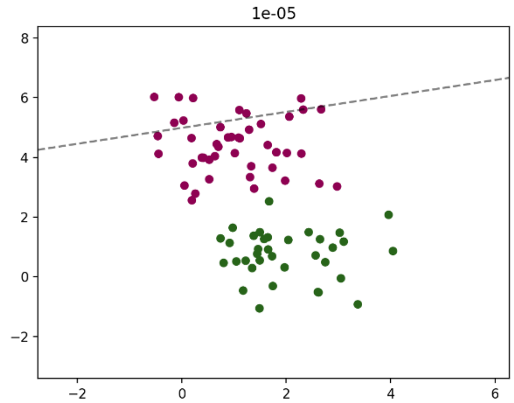
הפלט המתקבל –



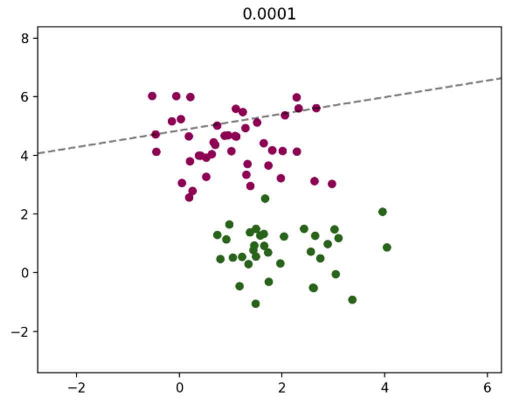
נבחין כי ערכי C הטובים ביותר מתקבלים עבור $C \geq 10^2 = 100$. עבור Cים כאלה, מקבלים אחוז דיוק של 100% על סט הואלידציה.
הפרמטר C מייצג את פרמטר הרגולזציה($regulazation$), כאשר ערכי C קטנים מאפשרים לנו להתעלם מאילוצים בצורה קלה יותר, מה שמאפשר אי-קיום האילוצים ותשלום של קנס קטן באופן יחסי על כך(שכן C קטן). כמו כן, ערכי C גדולים גורמים קנסות גדולים באופן יחסי על אי-קיום האילוצים של בעיית האופטימיזציה של SVM, ולכן ערכי C גדולים מאלצים יותר את קיום האילוצים בצורה הדוקה ומיטבית.
מקרה הקיצון בו $C = \infty$ מייצג את המקרה בו קיום כלל האילוצים הכרחי($corresponds\ to\ hard\ constraints$).

ה-$decision\ boundaries$ המתקבלים עבור ערכי ה-C השונים הם –

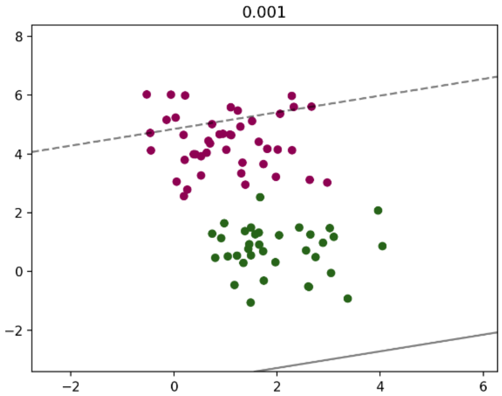
$C = 10^{-5}$



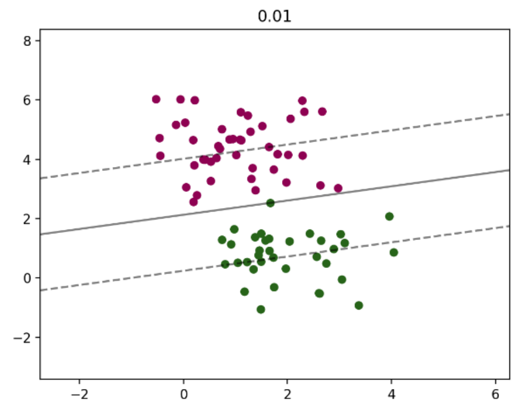
$C = 10^{-4}$



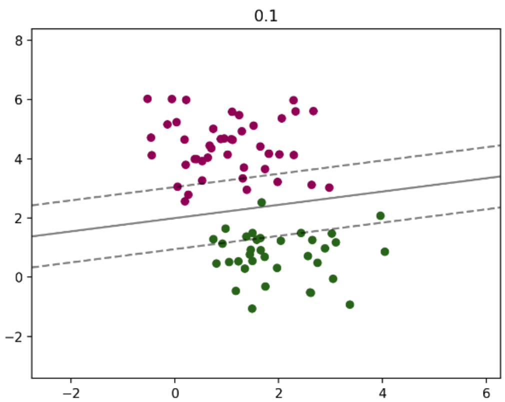
$C = 10^{-3}$



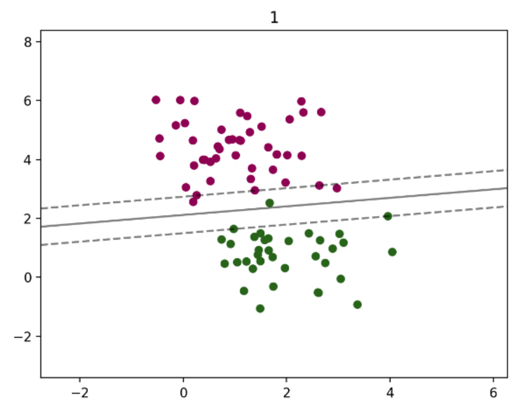
$C = 10^{-2}$



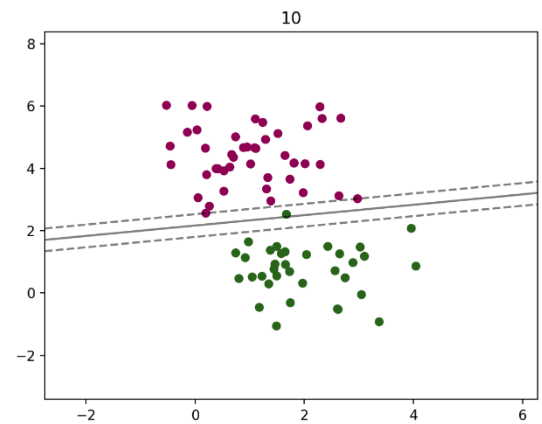
$C = 10^{-1}$



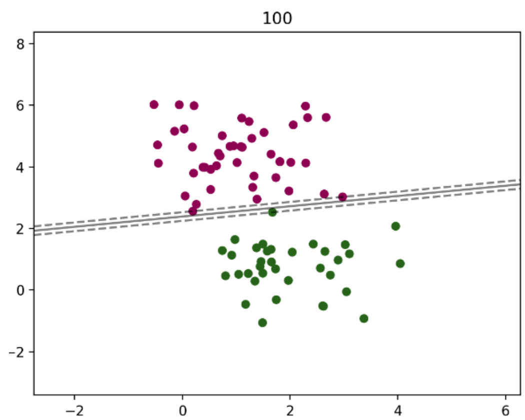
$C = 10^{0}$



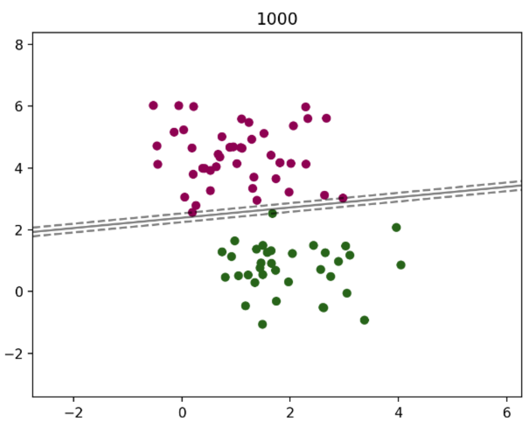
$C = 10^{1}$



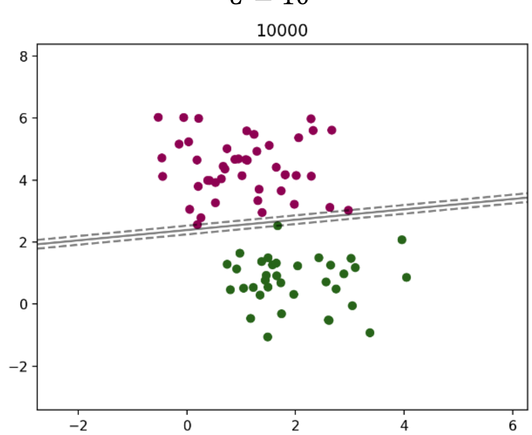
$C = 10^{2}$



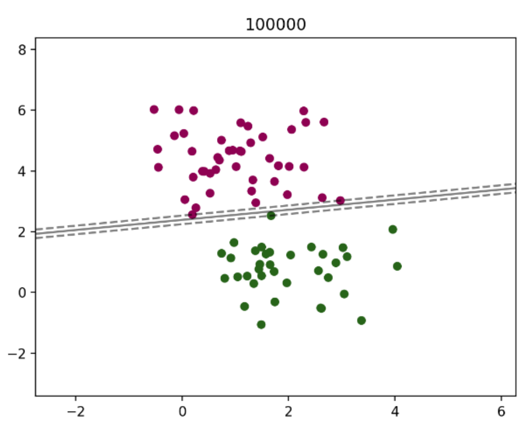
$C = 10^{3}$



$C = 10^{4}$



$C = 10^{5}$



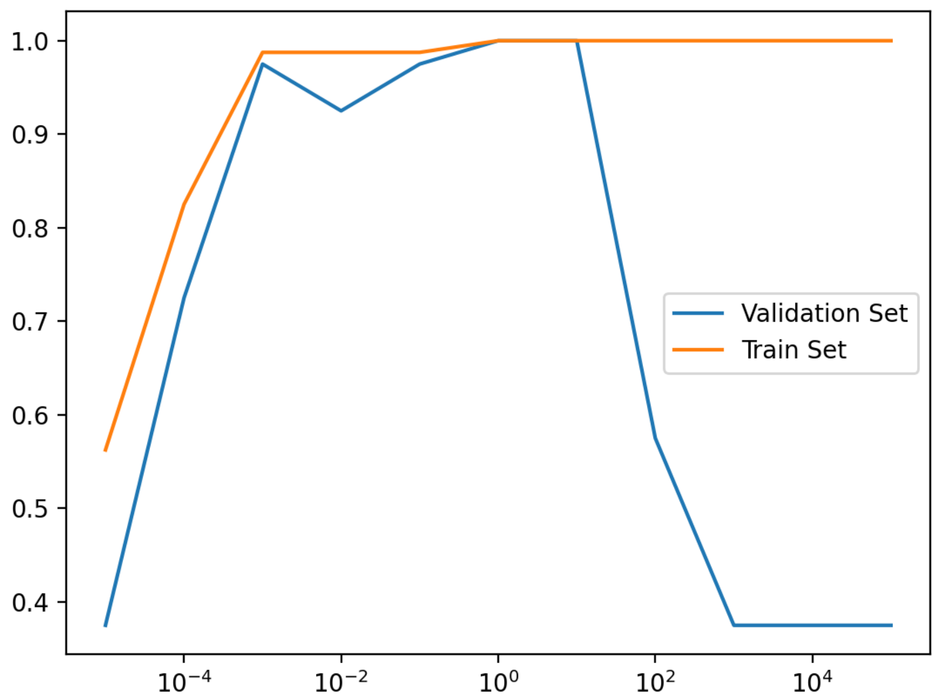
הסבר על הגרפים: ניתן לראות שככל שערכי הפרמטר C גדלים, אזי ה-$margins$ קטנים.
$margins$ גדולים יותר מאפשרים יותר טעויות, ונבחין שכאשר $C \geq 100$, אזי ה-$margins$ קטנים מספיק כך שלא יהיו טעויות כלל.
הפרמטר C מייצג את המדד לפיו עד כמה ניתן לסבול מטעויות בסיווג, וכאן ידוע כי הנתונים הם $separable$, ולכן ע"י בחירת C גדול מספיק(כאן $C \geq 100$), נקבל דיוק מושלם על סט הואלידציה.

סעיף ג. מימוש הפונק' $rbf\_accuracy\_per\_gamma$. בסעיף זה נבדוק מספר אפשרויות לערך הפרמטר γ ע"י אימון המודל עם ערכי $\gamma = \gamma_i$ שונים, ולאחר מכן בדיקת המודל על סט הואלידציה. נבחר ב-γ_i המביא את אחוז הדיוק המירבי על סט ה-$validation$.

הקוד –

```python
def rbf_accuracy_per_gamma(X_train, y_train, X_val, y_val):
    """
        Returns: np.ndarray of shape (11,) :
                    An array that contains the accuracy of the resulting model on the VALIDATION set.
    """
    valid_accuracy = np.zeros(11)
    train_accuracy = np.zeros(11)
    gammas = [pow(10, i) for i in range(-5, 6)]
    for i in range(11):
        clf = svm.SVC(C = 10, gamma=gammas[i])
        clf.fit(X_train, y_train)
        valid_accuracy[i] = calc_accuracy(X_val, y_val, clf)
        train_accuracy[i] = calc_accuracy(X_train, y_train, clf)

    valid_line, = plt.plot(gammas, valid_accuracy)
    train_line, = plt.plot(gammas, train_accuracy)
    plt.xscale('log')
    plt.legend((valid_line, train_line), ('Validation Set', 'Train Set'))
    plt.show()

    return valid_accuracy

```

הפלט המתקבל –



נבחין כי ערכי γ הטובים ביותר מתקבלים עבור $\gamma = 1, 10$. עבור γ-ים כאלה, מקבלים אחוז דיוק של 100% על סט הואלידציה.
כמו כן, ביצעתי נוספת של ערכי γ עם דיוק טוב יותר, והתוצאות עבור $\gamma \in \{1, 2, 3 \ldots, 10\}$ שהתקבלו היו די זהות, והשיגו דיוק של 100% על סט הואלידציה.

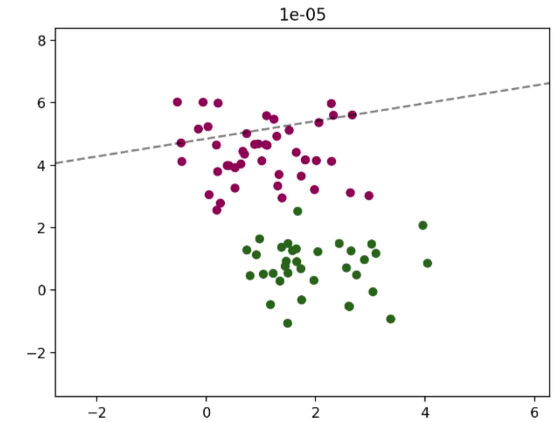
עבור $\gamma = 1$ ועבור $\gamma = 10$ מקבלים דיוק של 100% על סט הואלידציה, אך עבור פרמטר של $\gamma = 1$ מקבלים $decision\ boundaries$ רחבים יותר. כמו כן, נבחין שככל שהפרמטר γ גדל מקבלים $decision\ boundaries$ הדוקים יותר ויותר, ולכן בעצם מקבלים $overfitting$ של המודל על גבי הנתונים, ולפיכך מקבלים שגיאות גדולות יותר על גבי ה- $validation\ set$, מה שגורר ביצועים גרועים של המודל כאשר הפרמטר γ הינו גדול מדי.

מצד שני, כאשר הפרמטר γ קטן מדי, מקבלים $decision\ boundaries$ פחות ופחות הדוקים, ולכן בפרט מקבלים מודל בעל ביצועים פחות טובים שכן המודל בעל $margins$ גדולים יחסית, ולכן ייתכנו יותר שגיאות של המודל.

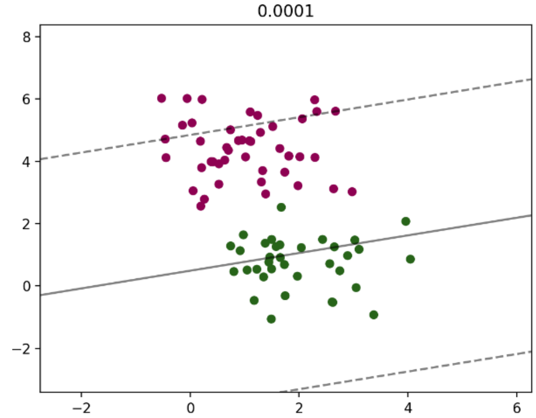
לפיכך, נסיק כי נרצה לבחור פרמטר γ סביר, כך שלא יהיה קטן מדי ויסבול משגיאות עקב $margins$ גדולים יחסית, וכן שלא יהיה גדול מדי ויסבול משגיאות עקב $overfitting$ של המודל על גבי נתוני סט האימונים.

ה-*decision boundaries* המתקבלים עבור ערכי ה-C השונים הם –

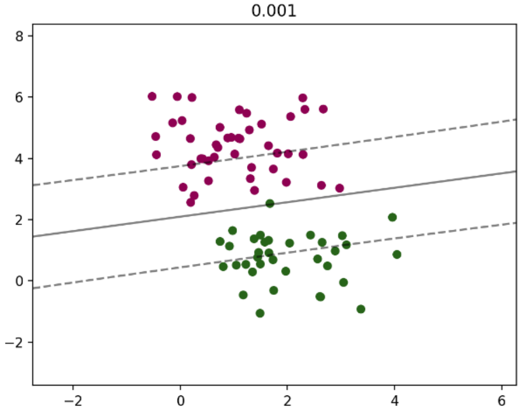
$\gamma = 10^{-5}$



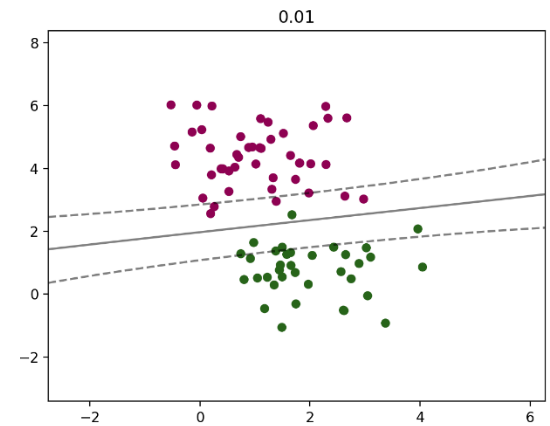
$\gamma = 10^{-4}$



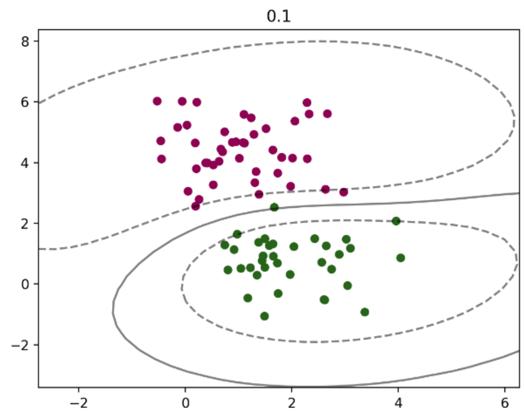
$\gamma = 10^{-3}$



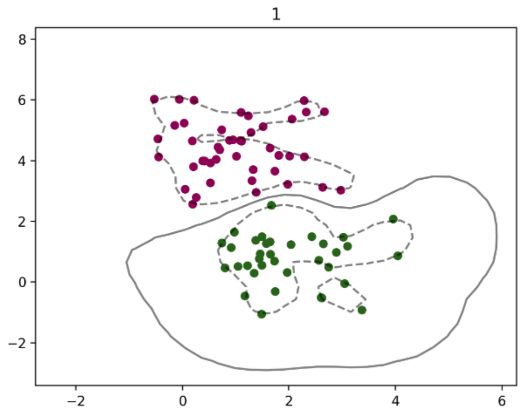
$\gamma = 10^{-2}$



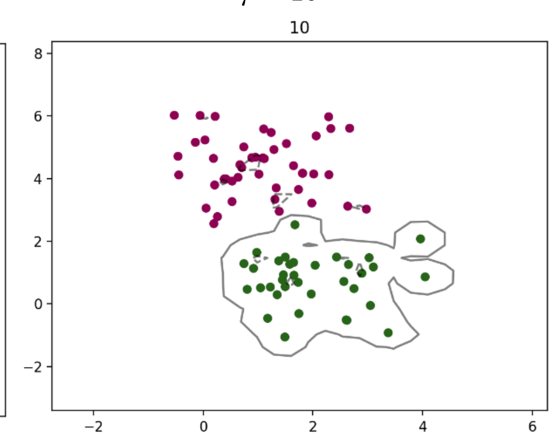
$\gamma = 10^{-1}$



$\gamma = 10^{0}$



$\gamma = 10^{1}$



$\gamma = 10^{2}$