

1. Let  $k(\mathbf{x}, \mathbf{x}')$  be a valid PSD kernel. Provide a valid PSD kernel  $\tilde{k}(\mathbf{x}, \mathbf{x}')$ , constructed from  $k$ , which is guaranteed to be normalized. That is, for all  $\mathbf{x}$  it holds that  $\tilde{k}(\mathbf{x}, \mathbf{x}) = 1$ . Prove your answer. (1)

PSD-kernel מוגדר כ- $\tilde{k} = \frac{k}{\sqrt{k(\mathbf{x}, \mathbf{x}) \cdot k(\mathbf{x}', \mathbf{x}')}}$ .  $\tilde{k}(\mathbf{x}, \mathbf{x}') := \frac{k(\mathbf{x}, \mathbf{x}')}{\sqrt{k(\mathbf{x}, \mathbf{x}) \cdot k(\mathbf{x}', \mathbf{x}')}}$  כ- $\tilde{k} : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$  מ- $\tilde{k}(\mathbf{x}, \mathbf{x}) = 1$

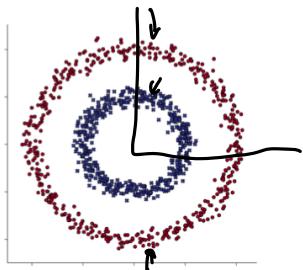
הוכיח נרמז. נסמן  $\tilde{k} = \frac{k}{\sqrt{k(\mathbf{x}, \mathbf{x}) \cdot k(\mathbf{x}', \mathbf{x}')}}$ .  $\tilde{k}(\mathbf{x}, \mathbf{x}') = \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\sqrt{k(\mathbf{x}, \mathbf{x}) \cdot k(\mathbf{x}', \mathbf{x}')}} = \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\|\psi(\mathbf{x})\| \cdot \|\psi(\mathbf{x}')\|}$ .

$$\sqrt{k(\mathbf{x}, \mathbf{x})} = \|\mathbf{x}\|$$

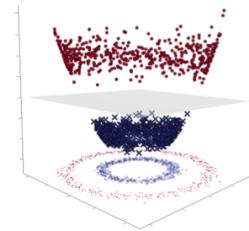
$\tilde{k}(\mathbf{x}, \mathbf{x}') = \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\|\psi(\mathbf{x})\| \cdot \|\psi(\mathbf{x}')\|} = \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\|\psi(\mathbf{x})\| \cdot \|\psi(\mathbf{x}')\|} = \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\|\psi(\mathbf{x})\| \cdot \|\psi(\mathbf{x}')\|}$

2. Consider a data set  $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  where  $\mathbf{x}_i \in \mathbb{R}^d$  and  $y_i \in \{\pm 1\}$ , and a feature map  $\psi : \mathbb{R}^d \rightarrow \mathcal{F}$  where  $\mathcal{F}$  is some feature space. Give an example of a data set  $S$  and a feature map  $\psi$  such that  $S$  is not linearly separable in  $\mathbb{R}^d$  (for  $d \geq 2$ ) but that the transformed data set  $S_\psi = \{(\psi(\mathbf{x}_i), y_i)\}_{i=1}^m$  is linearly separable in  $\mathcal{F}$ . (2)

הוכיח נרמז. על מנת ש- $\tilde{k}$  מוגדר כ- $\tilde{k}(\mathbf{x}, \mathbf{x}') = \frac{k(\mathbf{x}, \mathbf{x}')}{\sqrt{k(\mathbf{x}, \mathbf{x}) \cdot k(\mathbf{x}', \mathbf{x})}}$  מוגדר כ- $\tilde{k}(\mathbf{x}, \mathbf{x}) = 1$ .

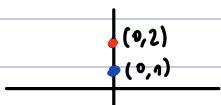


(a) Two class dataset that is not linearly separable



(b) Dataset mapped to  $\mathbb{R}^3$  using the mapping  $(x_1, x_2) \rightarrow (x_1^2, x_2^2, x_1^2 + x_2^2)^\top$

Figure 9.1: Illustration of mapping to Feature Space: Originally linearly inseparable dataset is linearly separable in mapped feature space. [Kernel Methods Examples](#)



$$S = \{(0,2), -1\}, \{(0,-2), -1\}, \{(0,1), 1\}$$

- רמה

: גורם גודל מוגדר כפונקציית סינגדון  $f(x,y) = \text{Sign}(w_2y + w_1x + w_0)$ . נאמר, אם  $f(x,y) > 0$ , אז  $w_2y + w_1x + w_0 > 0$ .

$$\left. \begin{array}{l} 2w_2 + w_0 < 0 \\ -2w_2 + w_0 < 0 \\ w_2 + w_0 > 0 \end{array} \right\} \iff -1 = \text{Sign}(2w_2 + w_0), -1 = \text{Sign}(-2w_2 + w_0), 1 = \text{Sign}(w_2 + w_0)$$

$$\Rightarrow -w_2 > \frac{w_0}{2} \wedge w_2 > \frac{w_0}{2} \wedge w_2 > -w_0$$

↓

$$w_2 - w_2 > -w_0 + \frac{w_0}{2} \wedge w_2 - w_2 > \frac{w_0}{2} + w_0 \Rightarrow 0 > w_0$$

$$0 < w_0 \iff 0 > -\frac{w_0}{2}$$

כגון עליי גואן. פה ניקי דג'ורי נרמזת כוונתנו בפונקציית סינגדון.

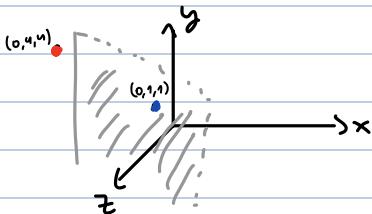
כל זוג זיהוי בפונקציית סינגדון יתאפשר על ידי הרכבתם.

$$\text{לדוגמא, } \Psi: \mathbb{R}^2 \rightarrow \mathbb{R}^3 \quad \Psi(x, y) = (x^2, y^2, x^2 - y^2)$$

$$\Psi(0,2) = (0,4,4), \Psi(0,-2) = (0,4,4), \Psi(0,1) = (0,1,1)$$

$$S_\Psi = \{\Psi(x_i), y_i\} = \{(0,4,4), -1\}, \{(0,4,4), -1\}, \{(0,1,1), 1\}$$

- רכיבי פונקציית העדרת סינגדון הם:



$$\begin{aligned} -z + 2 > 0 &\Leftrightarrow z < 2 \\ -1 < z < 2 &\text{ פול 1} \end{aligned} \quad : f(x) = \text{Sign}(-z + 2)$$

סימולר 1<2 נזקן, -1<2 פול 1, 1<2 נזקן, 2>2 פול 2, 1>2 נזקן, 2>2 פול 3.

3.  $k_1(\mathbf{x}, \mathbf{y})$  and  $k_2(\mathbf{x}, \mathbf{y})$  are valid kernels, then:

$$k_{\times}(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) \cdot k_2(\mathbf{x}, \mathbf{y})$$

הנ"ל כפלה קרטית  
kernel kernel

(3)

הנ"ל כפלה קרטית

is also a valid kernel. To prove this we'll use the fact that valid kernels are positive semi-definite.

You may find the following identities helpful (but don't have to use them):

$$\mathbf{x}^\top A \mathbf{y} = \text{Tr} [\mathbf{x}^\top A \mathbf{y}] = \text{Tr} [\mathbf{y} \mathbf{x}^\top A] \quad (1)$$

$$\text{Tr}[AB] = \sum_i [AB]_{ii} = \sum_i \sum_j A_{ij} B_{ji} \quad (2)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are vectors while  $A$  and  $B$  are matrices.

(a) Let  $k(\mathbf{x}, \mathbf{y})$  be a valid kernel and suppose that  $K$  is the kernel's Gram matrix over some finite set of points  $\{\mathbf{x}_i\}_{i=1}^N$ , such that  $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ . Show that for any finite set of  $N$  points, there exists some function  $f: \mathcal{X} \mapsto \mathbb{R}^N$  such that:

$$k(\mathbf{x}_i, \mathbf{x}_j) = f^\top(\mathbf{x}_i) f(\mathbf{x}_j) \quad (3)$$

where  $\mathcal{X}$  is space of the points  $\mathbf{x}_i$ . Using this fact, show that:

$$k_1(\mathbf{x}, \mathbf{y}) \cdot k_2(\mathbf{x}, \mathbf{y}) = \sum_i \sum_j g_i(\mathbf{x}) f_j(\mathbf{x}) f_j(\mathbf{y}) g_i(\mathbf{y}) \quad (4)$$

where  $k_1(\cdot, \cdot)$  and  $k_2(\cdot, \cdot)$  are valid kernels, and some functions  $f, g: \mathcal{X} \mapsto \mathbb{R}^N$ , where  $f_i(\mathbf{x})$  denotes the  $i^{th}$  index of the output of  $f(\mathbf{x})$ .

(b) Conclude that  $k_{\times}(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) \cdot k_2(\mathbf{x}, \mathbf{y}) = h^\top(\mathbf{x}) \cdot h(\mathbf{y})$  for some function  $h(\cdot)$ , thereby proving that  $k_{\times}(\cdot, \cdot)$  is a valid kernel.

הנ"ל כפלה קרטית

הנ"ל כפלה קרטית  
kernel kernel

הנ"ל כפלה קרטית

$$k_1(x, y) \cdot k_2(x, y) = f(x)^T f(y) \cdot g(x)^T g(y) = (\sum_j f_j(x) f_j(y)) (\sum_i g_i(x) g_i(y)) =$$

$$= \sum_i \sum_j f_j(x) \cdot f_j(y) g_i(x) \cdot g_i(y) = \sum_i \sum_j g_i(x) \cdot f_j(x) f_j(y) \cdot g_i(y)$$

הנחתה ש  $f, g$  סדרת פונקציות

- (b) Conclude that  $k_x(x, y) = k_1(x, y) \cdot k_2(x, y) = h^T(x) \cdot h(y)$  for some function  $h(\cdot)$ , thereby proving that  $k_x(\cdot, \cdot)$  is a valid kernel.

$$\forall x \in X \quad h(x) = \begin{pmatrix} g_1(x) f_1(x) \\ \vdots \\ g_1(x) f_N(x) \\ g_2(x) f_1(x) \\ \vdots \\ g_N(x) f_N(x) \end{pmatrix} \in \mathbb{R}^{N \times N}$$

: יישר מושג

$$h^T(x) h(y) = \sum_i h_i(x) \cdot h_i(y) = \sum_{(i,j) \in [N] \times [N]} g_i(x) \cdot f_i(x) f_j(y) \cdot g_j(y) = k_1(x, y) \cdot k_2(x, y) = k_x(x, y)$$

הנחתה ש  $\{x_i\}_{i=1}^N$  מושג  $X \subseteq \mathbb{R}^N$  ו  $f: X \rightarrow \mathbb{R}^{N \times N}$  היא כורט  $h: X \rightarrow \mathbb{R}^{N \times N}$  - כלומר  $K(x_i, x_j) = \langle h(x_i), h(x_j) \rangle$  (הנחתה ש  $\langle \cdot, \cdot \rangle$  מושג  $\mathbb{R}^{N \times N}$ )

.valid kernel      kernel       $K_x(\cdot, \cdot)$

## 1.2 PCA

Based on Lecture 9 and Recitation 11

4. Let  $X : \Omega \rightarrow \mathbb{R}^d$  be a random variable with zero mean and covariance  $\Sigma \in \mathbb{R}^{d \times d}$ . Show that for any  $v \in \mathbb{R}^d$ , where  $\|v\|_2 = 1$ , the variance of  $\langle v, X \rangle$  is not larger than variance obtained by the PCA embedding of  $X$  into a one-dimension subspace (assume that the PCA uses the actual  $\Sigma$ ).

$$\begin{aligned} \text{Var}(\langle v, X \rangle) &= E[(\langle v, X \rangle - E[\langle v, X \rangle])^2] = E[(v^t X - v^t E[X])^2] = E[(v^t X)^2] = \\ &= E[(v^t \cdot X)(v^t X)^t] = E[(v^t X)(X^t v)] = v^t E[X \cdot X^t] v = v^t \Sigma v \\ \Sigma &= E[(x - E[x])(x - E[x))^t] = E[X X^t] \quad E[X] = 0 \end{aligned}$$

גלו רצוי שקיים מילוי  $v$  ב  $\mathbb{R}^d$  כך ש  $v^t \Sigma v$  מינימלי, כלומר  $v_1 := \arg \max_{\|v\|_2=1} v^t \Sigma v$

לפ. 1 ניתן לראות שפונקציית PCA embedding יכולתPCA רצוייה. ס. ב

$$\text{Var}[\langle v, X \rangle] = v^t \Sigma v \leq v^t \Sigma v_1 = \text{Var}[\langle v_1, X \rangle]$$

### 1.3 Convex optimization

Based on Lecture 11 and Recitations 2,12

- Let  $f_1, \dots, f_m : C \rightarrow \mathbb{R}$  be a set of convex functions and  $\gamma_1, \dots, \gamma_m \in \mathbb{R}_+$ . Prove from definition that  $g(\mathbf{u}) = \sum_{i=1}^m \gamma_i f_i(\mathbf{u})$  is a convex function.
- Give a counterexample for the following claim: Given two functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$ , define a new function  $h : \mathbb{R} \rightarrow \mathbb{R}$  by  $h = f \circ g$ . If  $f$  and  $g$  are convex then  $h$  is convex as well.
- Let  $f : C \rightarrow \mathbb{R}$  be a function defined over a convex set  $C$ . Prove that  $f$  is convex iff its *epigraph* is a convex set, where  $\text{epi}(f) = \{(u, t) : f(u) \leq t\}$ .
- Let  $f_i : V \rightarrow \mathbb{R}$ ,  $i \in I$ . Let  $f : V \rightarrow \mathbb{R}$  given by

$$f(u) = \sup_{i \in I} f_i(u).$$

If  $f_i$  are convex for every  $i \in I$ , then  $f$  is also convex.

$$\begin{aligned} & \forall \gamma_1, \dots, \gamma_m \in \mathbb{R}_+ \quad \exists f_1, \dots, f_m : C \rightarrow \mathbb{R} \quad (1) \\ & \text{such that } \min_{u \in C} f_i - \ell \geq \gamma_i \quad \forall i \in [m] \quad \text{and} \quad \gamma_i \leq \gamma_j \quad \forall i, j \in [m] \\ & \gamma_i \in \mathbb{R}_+, \quad \forall i \in [m] \quad \text{and} \quad u, v \in C \\ & g(\alpha v + (1-\alpha)u) = \sum_{i=1}^m \gamma_i f_i(\alpha v + (1-\alpha)u) \leq \sum_{i=1}^m \gamma_i (\alpha f_i(v) + (1-\alpha) f_i(u)) = \alpha \sum_{i=1}^m \gamma_i f_i(v) + (1-\alpha) \sum_{i=1}^m \gamma_i f_i(u) = \\ & = \alpha g(v) + (1-\alpha) g(u) \end{aligned}$$

Ex 1:  $f(x) = -x$ ,  $g(x) = x^2$

$$f, g : \mathbb{R} \rightarrow \mathbb{R} \quad \forall x \in \mathbb{R} \quad f(x) = -x \quad g(x) = x^2 \quad (2)$$

$$\begin{aligned} & \text{such that } f''(x) = -1 < 0, \quad g''(x) = 2 \geq 0, \quad f'(x) = -1, \quad g'(x) = 2x \\ & f(-x) = -x, \quad g(-x) = x^2 \quad \text{and} \quad f(g(x)) = f(x^2) = -x^2 \quad \text{so} \quad h = f \circ g \end{aligned}$$

$$h''(x) = -2 < 0, \quad h'(x) = -2x \neq 0 \quad \text{so} \quad h \text{ is not convex.}$$

$$\begin{aligned} & \text{Definition: } \text{epi}(f) = \{(u, t) \mid f(u) \leq t\} \quad (3) \\ & \text{such that } \gamma_i \in \mathbb{R}_+, \quad \forall i \in [m] \quad \text{and} \quad \gamma_i \leq \gamma_j \quad \forall i, j \in [m] \end{aligned}$$

$$f(v) \leq f(u), \quad f(u) \leq f(w) \Rightarrow (v, f(v)), (u, f(u)), (w, f(w)) \in \text{epi}(f) \Rightarrow (\alpha v + (1-\alpha)u, \alpha f(v) + (1-\alpha)f(u)) \in \text{epi}(f)$$

$$\Downarrow \text{epi}(f) \text{ is convex}$$

$$f(\alpha v + (1-\alpha)u) \leq \alpha f(v) + (1-\alpha)f(u)$$

Ex 2:  $f(x) = x^2$

$f(u) \leq t_u, f(v) \leq t_v$   $\text{epi}(f) \cap \mathbb{R}^N \subseteq \{(u, t_u), (v, t_v)\} \in \text{epi}(f)$   $\Rightarrow$   $f$   $\text{convex}$   $\Leftrightarrow$   $\alpha u + (1-\alpha)v \in \text{epi}(f)$   $\forall \alpha \in [0, 1]$

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha)f(v) \leq \alpha t_u + (1-\alpha)t_v \implies (\alpha u + (1-\alpha)v, \alpha t_u + (1-\alpha)t_v) \in \text{epi}(f)$$

Definition of  $\text{epi}(f)$   $\vdash$

(4)

4. Let  $f_i : V \rightarrow \mathbb{R}$ ,  $i \in I$ . Let  $f : V \rightarrow \mathbb{R}$  given by

$$f(u) = \sup_{i \in I} f_i(u).$$

If  $f_i$  are convex for every  $i \in I$ , then  $f$  is also convex.

Definition of  $\text{Dom } f_i = V - \{u \in V \mid f_i(u) = -\infty, \forall i \in I\}$   $\text{Definir f}_i - \text{definir f}_i$   $\forall i \in I$   $\exists u \in V$   $\forall \alpha \in [0, 1]$

$$f(\alpha u + (1-\alpha)v) = \sup_{i \in I} f_i(\alpha u + (1-\alpha)v) \leq \sup_{i \in I} \alpha f_i(u) + (1-\alpha) f_i(v) = \alpha \sup_{i \in I} f_i(u) + (1-\alpha) \sup_{i \in I} f_i(v) = \alpha f(u) + (1-\alpha) f(v)$$

Proof of  $f$  convex:  $\forall u, v \in V$   $\forall \alpha \in [0, 1]$   $\sup(A+B) = \sup(A) + \sup(B)$   $\sup(\alpha A) = \alpha \sup(A)$

$$= \alpha f(u) + (1-\alpha) f(v)$$

## 1.4 Sub-gradients for Soft-SVM Objective

Based on Lecture 11 and Recitations 2,12

The Soft-SVM objective, though convex, is not differentiable in all of its domain due to the use of the hinge-loss. Therefore, to implement a sub-gradient descent solver for this problem we must first describe sub-gradients of the objective.

5. Given  $\mathbf{x} \in \mathbb{R}^d$  and  $y \in \{\pm 1\}$ . Show that the hinge loss is convex in  $\mathbf{w}, b$ . That is, define

$$\text{הו נגיף } \ell_{\mathbf{x},y}^{\text{hinge}} \rightarrow f(\mathbf{w}, b) := \ell_{\mathbf{x},y}^{\text{hinge}}(\mathbf{w}, b) = \max(0, 1 - y(\mathbf{x}^\top \mathbf{w} + b))$$

and show that  $f$  is convex in  $\mathbf{w}, b$ .

6. Deduce some sub-gradient of the hinge loss function  $g \in \partial \ell_{\mathbf{x},y}^{\text{hinge}}(\mathbf{w}, b)$ .

7. Let  $f_1, \dots, f_m : \mathbb{R}^d \rightarrow \mathbb{R}$  be a set of convex functions and  $\mathbf{g}_k \in \partial f_k(\mathbf{x})$  for all  $k \in [m]$  be sub-gradients of these functions. Define  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  by  $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$ . Show that  $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$ .

8. Let  $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{\pm 1\}$  be a sample and define  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  by:

$$f(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m \ell_{\mathbf{x}_i, y_i}^{\text{hinge}}(\mathbf{w}, b) + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Find a sub-gradient of  $f$  for any  $\mathbf{w}$ .

$$\forall (\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} \quad f_2(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b), \quad f_1(\mathbf{w}, b) = 0 \quad \text{רנפ' } f_1, f_2 : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R} \quad \text{רנפ' } f_1, f_2 \text{ הינה כונין}$$

$$\nabla f_1(\mathbf{w}, b) = (0, 0), \quad \nabla f_2(\mathbf{w}, b) = \nabla (1 - y \mathbf{x}^\top, 1)(\mathbf{w}, b) = (0, 0) \quad \text{רנפ' } f_1, f_2 \text{ הינה כונין}$$

סע נאנו בדינור גפ' הרכבת הינה (הניעו פה ורמזו)  $f_1, f_2$  הינה כונין.

4. Let  $f_i : V \rightarrow \mathbb{R}$ ,  $i \in I$ . Let  $f : V \rightarrow \mathbb{R}$  given by

$$f(u) = \sup_{i \in I} f_i(u).$$

If  $f_i$  are convex for every  $i \in I$ , then  $f$  is also convex.

$$(f_i \text{ הינה כונין } \Rightarrow f_i \text{ הינה כונין }) \Rightarrow f \text{ הינה כונין } f(\mathbf{w}, b) = \max(f_1, f_2) \quad \text{רנפ' } f_1, f_2 \text{ הינה כונין}$$

$$\ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) \text{ הינה subgradient} \quad \text{רנפ' } f_1, f_2 \text{ הינה כונין} \quad (6)$$

$$\text{הנ' } f_1 \text{ הינה כונין } \Rightarrow f_1(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b) \quad \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) < 1 \quad \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) \geq 1$$

$$\text{הנ' } f_2 \text{ הינה כונין } \Rightarrow f_2(\mathbf{w}, b) = 0 \quad \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) \geq 1$$

$$\ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = \begin{cases} 0 & \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) \geq 1 \\ \infty & \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) < 1 \end{cases} \quad \text{רנפ' } f_1, f_2 \text{ הינה כונין}$$

$$y \in \{-1, 1\} \quad \text{רנפ' } y \neq 0 \quad \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) < 1 \quad \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) \geq 1$$

$$\ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = \begin{cases} 0 & \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) \geq 1 \\ \infty & \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) < 1 \end{cases} \quad \text{רנפ' } f_1, f_2 \text{ הינה כונין}$$

$$V = \begin{cases} 0 & \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) \geq 1 \\ -\infty & \text{רנפ' } y(\mathbf{x}^\top \mathbf{w} + b) < 1 \end{cases} \quad \text{רנפ' } f_1, f_2 \text{ הינה כונין}$$

7. Let  $f_1, \dots, f_m : \mathbb{R}^d \rightarrow \mathbb{R}$  be a set of convex functions and  $\mathbf{g}_k \in \partial f_k(\mathbf{x})$  for all  $k \in [m]$  be sub-gradients of these functions. Define  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  by  $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$ . Show that  $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$ . (7)

: פ"גMN sk,  $x \in \mathbb{R}^d$  -> ANKLNW ג"א פ"ג. ג"אLNW נKO  $g_1, \dots, g_k - 1$ , LNWLDP  $f_1, \dots, f_m$  -> NCIL Q

$$\forall u \in \mathbb{R}^d \quad \forall i \in [m] \quad f_i(u) \geq f_i(x) + \langle g_i, u - x \rangle$$

$$\sum_{i=1}^m f_i(u) \geq \sum_{i=1}^m f_i(x) + \sum_{i=1}^m \langle g_i, u - x \rangle = \langle \sum_{i=1}^m g_i, u - x \rangle$$

: פ"גMN f:  $\mathbb{R}^d \rightarrow \mathbb{R}$  st  $\forall x \in \mathbb{R}^d \quad f(x) = \sum_{i=1}^m f_i(x)$  LNWLDP ג"אLNW |

$$\forall u \in \mathbb{R}^d \quad \forall i \in [m] \quad f(u) \geq f(x) + \langle \sum_{i=1}^m g_i, u - x \rangle$$

$$\sum_{i=1}^m g_i \in \partial f = \partial \sum_{i=1}^m f_i \quad \text{פ"גMN ג"אLNW נKO ג"אLNW |}$$

8. Let  $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{\pm 1\}$  be a sample and define  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  by: (8)

$$f(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m \ell_{\mathbf{x}_i, y_i}^{hinge}(\mathbf{w}, b) + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Find a sub-gradient of  $f$  for any  $\mathbf{w}$ .

$$\begin{aligned} \mathbf{g} &:= \begin{cases} (\overset{\circ}{0}) & y_i(x_i^\top \mathbf{w} + b) \geq 1 \\ (-y_i x_i) & y_i(x_i^\top \mathbf{w} + b) < 1 \end{cases} \\ &\text{לפ"גLNW נKO 1je' } \ell_{\mathbf{x}_i, y_i}^{hinge}(\mathbf{w}, b) \text{ LNWLDP 1je' } \end{aligned}$$

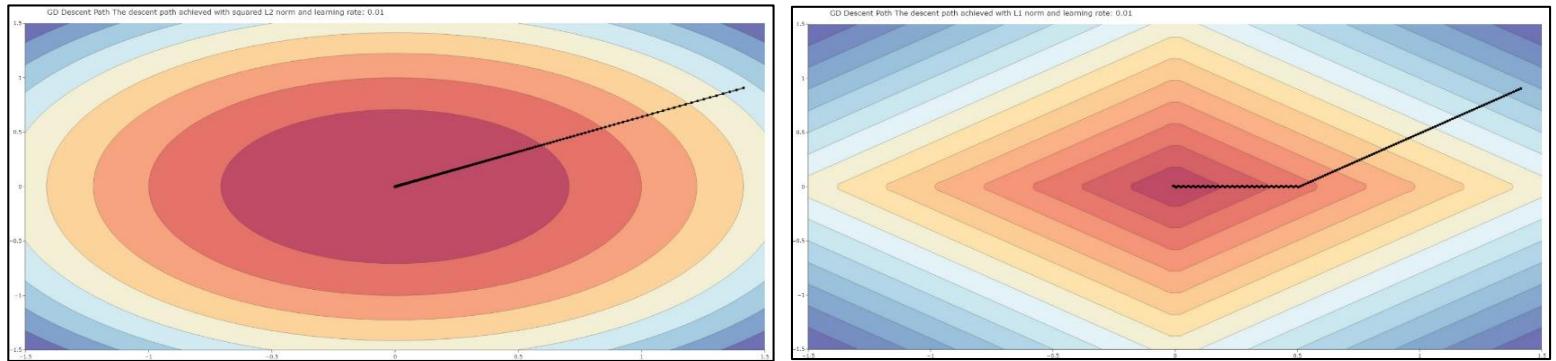
לפ"גLNW נKO 1je'  $\sum_{i=1}^m g_i \in \partial f$   $\Leftrightarrow$   $\sum_{i=1}^m g_i \in \partial \left( \frac{1}{m} \sum_{i=1}^m \ell_{\mathbf{x}_i, y_i}^{hinge}(\mathbf{w}, b) \right)$

$$\hat{f}'(\mathbf{w}, b) = \{ \nabla \hat{f}(\mathbf{w}, b) \} = \left\{ \begin{pmatrix} \lambda \mathbf{w} \\ 0 \end{pmatrix} \right\} \in \mathbb{R}^{d+1} \cap \mathbb{R} \quad \text{LNWLDP ג"אLNW} \quad \text{לפ"גLNW} \quad \hat{f} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R} \quad \hat{f} := \frac{\lambda}{2} \|\mathbf{w}\|^2 + b$$

$$\frac{1}{m} \sum_{i=1}^m g_i + \begin{pmatrix} \lambda \mathbf{w} \\ 0 \end{pmatrix} \in \partial f(\mathbf{w}, b) \quad \text{לפ"גLNW} \quad \text{לפ"גLNW} \quad \text{לפ"גLNW}$$

## חלק פרקי IML תרגיל 5:

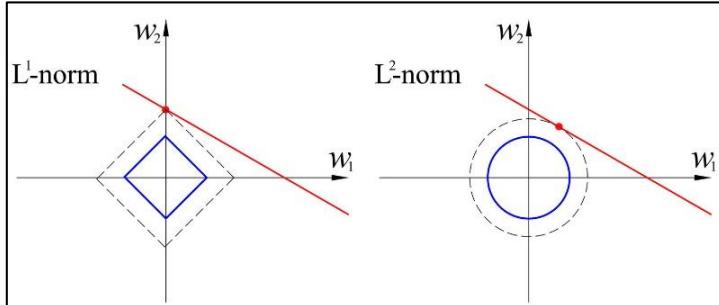
**1. Plot the descent path for each of the settings described above (you can use the `plot_descent_path`). Add below the plots for  $\eta = 0.01$  and explain the differences seen between the L1 and L2 modules.**



ברקע של הגרפים ניתן לראות את השמי בערבי פונקציית objective (L1 או L2) בכל צדי העדכו של אלגוריתם - gradient descent בחיפוש אחר הערכים  $w_1, w_2$ , אשר יובילו לנקודת המינימום של הפונקציה, כאשר כחול מייצג ערכים גבוהים ואדום מייצג ערכים נמוכים.

ישנם שני הבדלים העיקריים בין המודלים:

- בrukע של הגרפים, ניתן לראות שעבור L1 גבולות ערכי הפונקציה "ריבועיים" בעוד שעבור L2 (בריבוע) הם "עוגליים", בהתאם לפונקציות:



- הבדל נוסף הוא, שבعود ש-L2 מתקדמת ישירות אל נקודת המינימום  $(0,0)$ , L1 מתקדמת ראשית אל הנקודה  $(0,0.5)$ , ולאחר מכן מתחילה לנוע שמאלה אל עבר נקודת המינימום  $(0,0)$ . זה מתאים לכך ש-L1 מכובצת את המשקלות - L1 מעודדת דילולות ואייפוס של פיצרים (L1 קודם כל מאפסת פיצ'ר בלשחו בניסיון להקטין, לעומת זאת L2 בכל הדרך מכובצת את כל הפיצרים באופן דומה).
- בנוסף ניתן לראות שעבור L2 בתחלתו ישנים עצדים גדולים, ובמשך תקופה מסוימת מינימים העדדים נהים קטנים יותר (על אף שקצב הלמידה נמוך). זה מटבעא באופי הנטיבת הנקודות על גבי הנטיב. לעומת זאת עבור L1, צעדיו העדכו הם באותו גודל לאורך הנטיב (נקודות על סמרק ההיפר פרטמר), ומה שמשמעותה זה רק הסימן, כי הנגדות היא  $+1$  או  $0$  במקרה של סאב הגרידינט-ב-0).

נשים לב ש- $w$ , עבור שתי הפונקציות, מסיים קרוב לראשית היצירים, ל- $(0,0)$ , שהוא אבן, מינימום (גלובלי) עבורן. שתי הפונקציות קמורות, ולכן מינימום מקומי = מינימום גלובלי, ולכן הגיוני שהצלחנו להגיע בקרוב לערכיהם הנתונים מינימום.

**2. Describe two phenomena that can be seen in the descent path of the  $\ell_1$  objective when using GD and a fixed learning rate**

אתהาร שתי תופעות מעניינות שניתן לזרות לגבי נתיב הירידה של 1 המוצג בגרף:

- הגרדיאנט משנה את ביומו במהלך הדרכ (כלומר הוא לא יציב לאורק הנתיב כמו במרקה של 2)

- ישן תנודות של 2 א סביר היישר 0=y החול משלב מסוים

אנמק את התופעות:

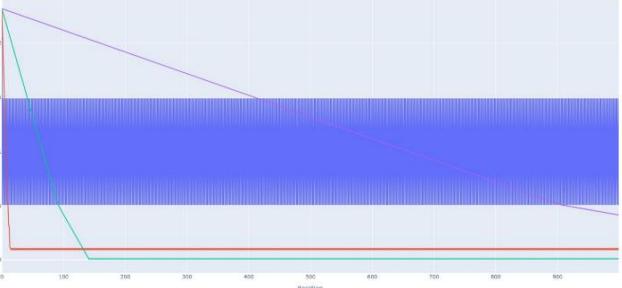
- ניתן לראות שהכיוון אליו מצביע הגרדיאנט בנקודות הפתיחה (בחלק הכחול) אינו נקודת המינימום (0,0),

ולכן בתניב ההתקדמות של 1 מתקדים באופן מסוים (עם גראדיאנט 1,1) לכל אורק הדרכ, עד שמגעים לאחור האופטימי עברו 2 א (אפס), ובשלב זה משנים את צורת ההתקדמות, בניסיון להביא את 1 א לאחור של 0 גם כן.

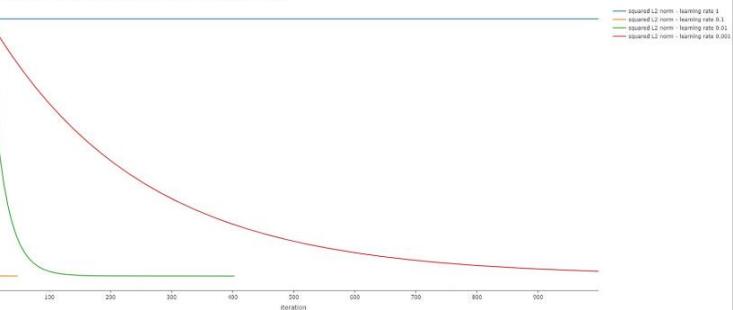
- החול משלב זה, הנגזרת החלקית של 2 א משנה סימן לסירוגין (מאחד למינוס אחד ולהפך), מה שמתבטא בתנודות קטנות סביר היישר 0=y. ניתן להסביר זאת על ידי כך שהשלב בו 0~2 א, קצב הלימוד מהיר מדי, ומקשה על התכנסות הפונקציה למקולת המשיפה מינימום (0=2 א). במקביל לניסיון של 2 א להתכנס ל-0, ממשיכים ליעיל את 1 א, עד שלבסוף מגעים קרוב מספיק לראשית היצרים ונעצרים.

### 3. For each of the modules, plot the convergence rate (i.e., the norm as a function of the GD iteration) for all specified learning rates. Explain your results.

Convergence rate for L1 norm function for different learning rates



Convergence rate for squared L2 norm function for different learning rates



ניתן לראות את השינוי בתוכניות למינימום

כפונקציה של קצב הלימוד.

עבור קצב למידה גבוה במיוחד יש "overshoot", יש פלקטוואציות סביר המינימום בר שלא מצליחה להוועץ התוכניות למינימום.

עבור קצב למידה נמוך במיוחד התוכניות איטיתים, ולא מצליחים למצוא את המינימום, על אף שמתקדמיים בכיוון הנכון.

גם עבור קצבי הלמידה הבינוויים ניתן לראות שמתקיים טריז אוף בין קצב התוכניות לבין פלקטוואציות (ש망tauאות ב"התקנות" על אותו ערך של הפונקציה ואי הצלחה לרדת בעור הפונקציה) שמנועות הגעה למינימום האמיתי.

נבחן שובו ש- 2 א בקצבו הלימוד המאוזנים (לא קטנים מדי ולא גדולים מדי), מצליחים לענות על תנאי העזירה של שינוי קטן מספיק ולכן נעצרים, ב- 1 א ממשיכים לנסות להתייעל עד למספר עדכוניים מקסימלי (חסם מספר האיטרציות), בולם, לעומת מגעים למינימום לגמרי מספק.

זה נובע מכך שב-2 א העדכון מושפע גם מהגודל של א וב-1 א, ולכן יש קשיי גדול יותר להתכנס (המינימום הוא באפס ולכן בשא קרוב לאפס קצב הלימוד קטן וכך קל יותר להתכנס בדיק לנקודת נקודה ולא לעשות פלקטוואציות סביבה).

הגרפים הללו מדגימים את החשיבות של בחירת קצב לימוד המאזן את הצורך להתכנס בזמן סביר, ולא לבצע פלקטוואציות סביר המינימום ובכך "לפספס" אותו ולא להצליח לשפר את המשקولات.

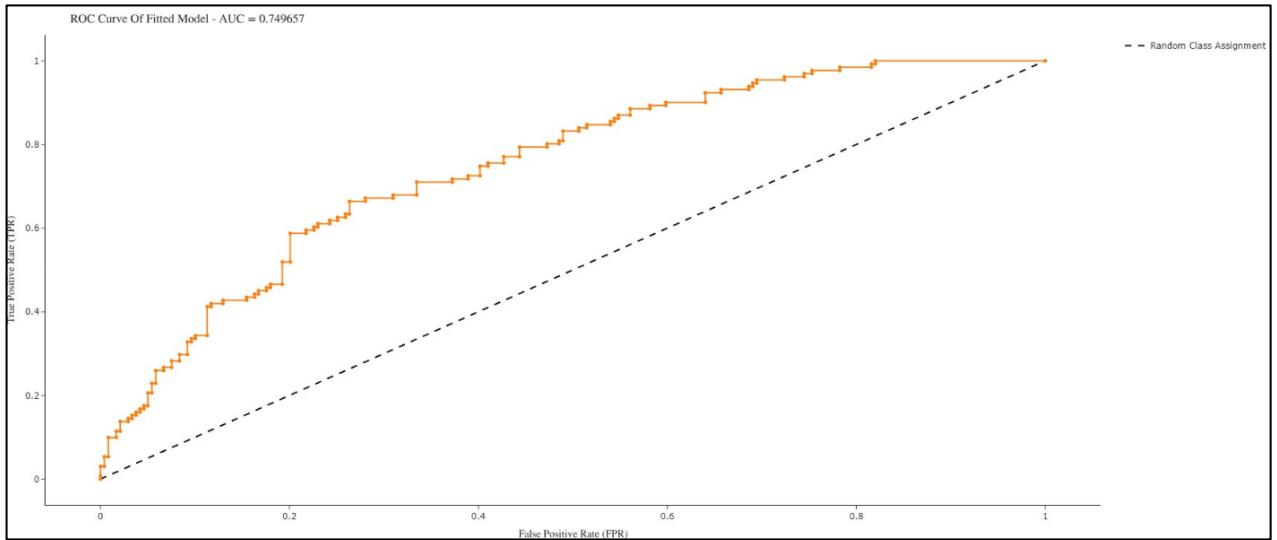
### 4. What is the lowest loss achieved when minimizing each of the modules? Explain the differences.

Module L1 norm achieves lowest loss: 0.00812 with learning rate: 0.01

Module squared L2 norm achieves lowest loss: 0.0 with learning rate: 0.1

כאשר עיגלתי את התוצאות חמיש ספירות אחרי הנקודה. אנחנו רואים שהקצב לימוד האופטימלי עבר ל-2 (בריבוע) הוא גדול יותר, ושהיא מצליחה להגיע למינימום הגלובלי, בעוד ש-1 מושג קרוב. נוכל להסביר זאת כר-ב-2, צעד העדכון לוחק בכוון של הדרך הקצרה ביותר אל המינימום. הגרדיאנט עצמו הוא  $2^*w$ , ולכן ככל ש-w גדול יותר, כך צעד העדכון גדול יותר, וכך אף אפקט אשר דואג להקטנה של קצב הלמידה ונהיים קטנים יותר. לכן, על אף קצב הלמידה הקבוע, ישנו עוד אפקט אשר דואג להקטנה של קצב הלמידה באזוריים קריטיים שעוזר לתנוודות ולהתכנס, גם עם קצב לימוד מעט גדול. לעומת זאת, עבור L1, כפי שראינו בגרף בסעיף 1, מתקבלות תנודות סביב הישר  $0 = y$ , המדגימות שנדרש קצב לימוד קבוע נמוך יחסית כדי לאפשר צעדן קטןים שלא יהוו overshoot בשמתקבים למינימום ויקשו על הגעה למינימום. ההבדל בהצלחה למצאו את המינימום נובע מכך ש-1 פונקציה שאיננה גדרה בכל התחומים (או גדרות ב-0), וכן w מגע לערך זהה, לאחר מכן משתמשים בסא-גרדיאנט במקום גרדיאנט, אנחנו רואים שינוי במוגמת העדכנים), ומופיע הדليلות של L1 אשר מנסה לעשות feature selection.

#### 8. Using your implementation, fit a logistic regression model over the data. Use the predict\_proba to plot an ROC curve



#### 9. Which value of $\alpha$ achieves the optimal ROC value according to the criterion below. Using this value of $\alpha$ \* what is the model's test error? $\alpha^* = \operatorname{argmax}_\alpha \{TPR_\alpha - FPR_\alpha\}$

The optimal threshold for TP & FP ratios is: 0.97135. using this threshold we achieve test error: 0.31522

כאשר עיגלתי חמיש ספירות אחרי הנקודה.

#### 10. What value of $\lambda$ was selected and what is the model's test error? (L1)

selected lambda for l1 regularization term is: 0.02, getting test error of: 0.25

#### 11. What value of $\lambda$ was selected and what is the model's test error? (L2)

selected lambda for l2 regularization term is: 0.001, getting test error of: 0.28261

כאשר עיגלתי חמיש ספירות אחרי הנקודה.

