NAME: GEORGIOS DIMITROPOULOS

STUDENT NUMBER: 4727657

EMAIL ADDRESS: G.Dimitropoulos1@student.tudelft.nl

Machine Learning: Assignment 1
Regularization and Sparsity

$\boxed{\text{Assignment 01}}$

01 a

$$\begin{cases} x_1 \to y_1 \to + = -1 \\ x_2 \to y_1 \to + = -1 \end{cases}$$

$\gamma = 1$

$r_- = 1$

$r_+, r_- \in \mathbb{R}$

$\boxed{\bullet\ \lambda = 0}$: $L(r_+) = \sum_{i=1}^{2} \frac{1}{N_{y_i}} \| x_i - r_{y_i} \|$

$= \frac{1}{N_{y_1}} \left[\| x_1 - r_+ \| + \| x_2 - r_+ \| \right] +$

$N_{y_1} = N_{r_+} = 2$

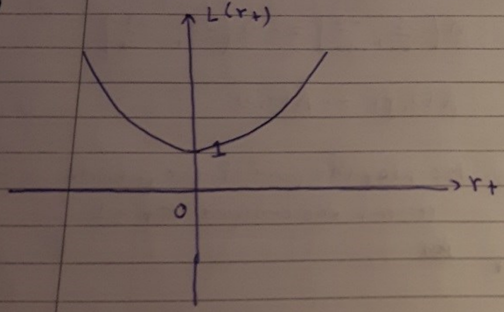
$= \frac{1}{2} \| x_1 - r_+ \| + \frac{1}{2} \| x_2 - r_+ \| = \frac{1}{2} \| 1 - r_+ \| + \frac{1}{2} \| -1 - r_+ \|$

<u>We know that</u>. $\| q - p \|^2 = \sqrt{(q-p)(q-p)}^2 = |(q-p)|^2 = (q-p)^2$

So, $L(r_+) = \frac{1}{2} (1 - r_+)^2 + \frac{1}{2} (-1 - r_+)^2 = \frac{1}{2} (1 - 2r_+ + r_+^2) + \frac{1}{2} (1 + 2r_+ + r_+^2)$

$= \frac{1}{2} - r_+ + \frac{r_+^2}{2} + \frac{1}{2} + r_+ + \frac{r_+^2}{2} = r_+^2 + 1$

So, $\boxed{L(r_+) = r_+^2 + 1}$ (parabola)



$\boxed{\bullet\ \lambda = 1}$ $L(r_+) = r_+^2 + 1 + \| 1 - r_+ \|_1 = r_+^2 + 1 + |1 - r_+|$

(if dimension $= 1 \Rightarrow \| a - b \|_1 = |a - b|$)

We distinguish <u>2 cases</u>:

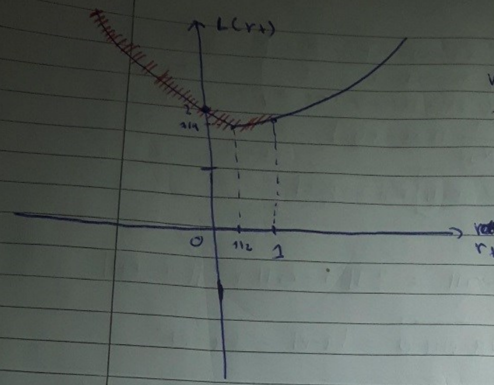
A). $1 - r_+ \geq 0 \Rightarrow \boxed{r_+ \leq 1}$, then $|1 - r_+| = 1 - r_+$

So, $L(r_+) = r_+^2 + 1 + 1 - r_+ = \boxed{r_+^2 - r_+ + 2}$, $\Delta = 1 - 4 \cdot 2 = -7 < 0$

↳ <u>polynomial</u> So, $\boxed{L(r_+) = r_+^2 - r_+ + 2}$

$M\left(-\frac{b}{2a}, -\frac{\Delta}{4a}\right) \Rightarrow M\left(\frac{1}{2}, \frac{7}{4}\right)$

$A(0, \gamma) \Rightarrow A(0, 2)$



We keep the part of the parabola that respects the constraint $r_t \leq 1$ (red color)

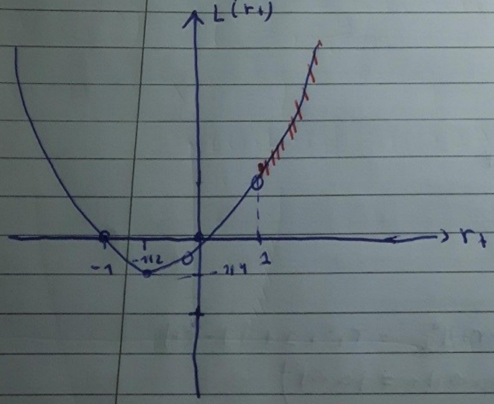
B) $\cdot\ 1 - r_t < 0 \Rightarrow \boxed{r_t > 1}$ then $|1 - r_t| = -1 + r_t$

So, $\boxed{L(r_t)} = r_t^2 + 1 - 1 + r_t = \boxed{r_t^2 + r_t}$.

$r_t^2 + r_t = 0 \Leftrightarrow r_t(r_t + 1) = 0 \Leftrightarrow r_t = 0$ or $r_t = -1$, $\Delta = 1$

$M\left(\frac{-b}{2a}, \frac{-\Delta}{4a}\right) \Rightarrow M\left(-\frac{1}{2}, -\frac{1}{4}\right)$

$A(0, \gamma) \Rightarrow A(0, 0)$



We keep the part of the parabola that respects the constraint $r_t > 1$

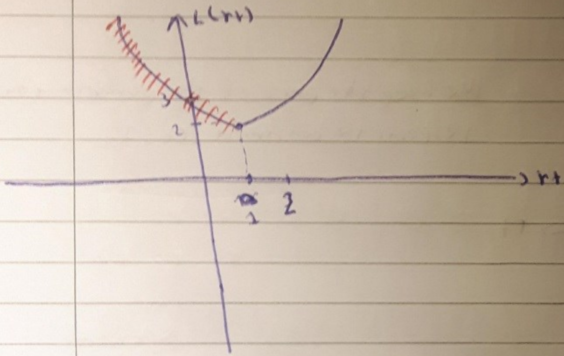
$\boxed{\lambda = 2}$ $L(r_t) = r_t^2 + 1 + 2|1-r_t|$

A). $1 - r_t \geq 0 \Rightarrow \boxed{r_t \leq 1}$ then $L(r_t) = r_t^2 + 1 + 2 - 2r_t = \boxed{r_t^2 - 2r_t + 3}$

$\Delta = 4 - 12 = -8$

$M\left(\frac{-b}{2a}, \frac{-\Delta}{4a}\right) \Rightarrow M(1, 2)$

$A(0, \gamma) \Rightarrow A(0, 3)$

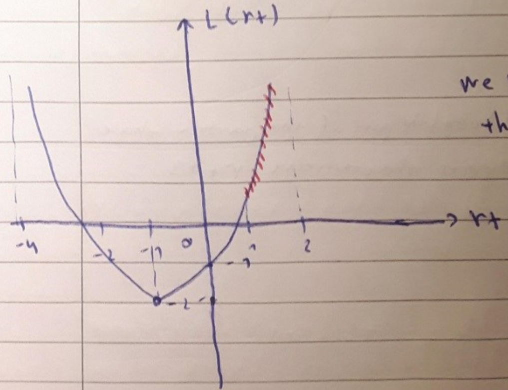


we keep the part of the parabola that respects the constraint $r_t \leq 1$.

B). $1 - r_t < 0 \Rightarrow \boxed{r_t > 1}$ then $L(r_t) = r_t^2 + 1 - 2 + 2r_t = \boxed{r_t^2 + 2r_t - 1}$

$\Delta = 4 - 4 \cdot 1 \cdot (-1) = 8$

$M\left(\frac{-b}{2a}, -\frac{\Delta}{4a}\right) \Rightarrow M(-1, -2)$ $A(0, \gamma) \Rightarrow A(0, -1)$



we keep the part of the parabola that respects the constraint $r_t > 1$

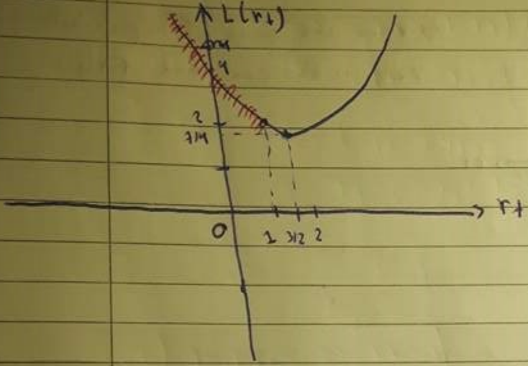
$\boxed{\cdot \lambda = 3}$ $\quad L(r_t) = r_t^2 + 1 + 3|1 - r_t|$

A). $1 - r_t \geq 0 \Rightarrow \boxed{r_t \leq 1}$ then $\boxed{L(r_t)} = r_t^2 + 1 + 3 - 3r_t$

$= \boxed{r_t^2 - 3r_t + 4}$

$\Delta = 9 - 4 \cdot 4 = -7$

$M\left(\frac{-b}{2a}, \frac{-\Delta}{4a}\right) \Rightarrow M\left(\frac{3}{2}, \frac{7}{4}\right)$ $\quad A(0, \gamma) \Rightarrow A(0, 4)$



We keep the part of the parabola that respects the constraint $r_t \leq 1$.

.B) $\cdot$ $1 - r_t < 0 \Rightarrow \boxed{r_t > 1}$ then $\boxed{L(r_t)} = r_t^2 + 1 - 3 + 3r_t$

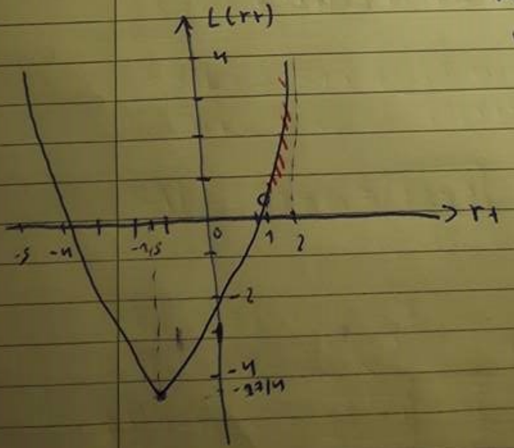
$= \boxed{r_t^2 + 3r_t - 2}$

$\Delta = 9 - 4 \cdot (-2) = 17$

$M\left(\frac{-b}{2a}, \frac{-\Delta}{4a}\right) \Rightarrow M\left(\frac{-3}{2}, \frac{-17}{4}\right)$

$A(0, \gamma) \Rightarrow A(0, -2)$

We keep the part of the parabola that respects the constraint $r_t > 1$.



(b) We are going to find the derivative of $L(r_+)$ with respect to r_+.

$\boxed{\lambda = 0}$ $\quad L(r_+) = r_+^2 + 1$

$$\frac{\partial L(r_+)}{\partial r_+} = 0 \Leftrightarrow (r_+^2 + 1)' = 0 \Leftrightarrow 2r_+ = 0 \Leftrightarrow \boxed{r_+ = 0}$$

$\boxed{\lambda = 1}$ $\quad L(r_+) = r_+^2 + 1 + |1 - r_+|$

$$\frac{\partial L(r_+)}{\partial r_+} = 2r_+ + |1-r_+|' = 2r_+ + \frac{1-r_+}{|1-r_+|}\cdot(-1) =$$

$$= 2r_+ + \frac{r_+ - 1}{|1-r_+|}$$ It should be $1 - r_+ \neq 0 \Rightarrow \boxed{r_+ \neq 1}$

So $$\frac{\partial L(r_+)}{\partial r_+} = 0 \Leftrightarrow \frac{2r_+\cdot|1-r_+|}{\cancel{|1-r_+|}} + \frac{r_+ - 1}{\cancel{|1-r_+|}} = 0$$

We distinguish again 2 cases:

A). $1 - r_+ > 0 \Rightarrow \boxed{r_+ < 1}$ then $|1-r_+| = 1 - r_+$

So $$\frac{\partial L}{\partial r_+} = 0 \Leftrightarrow 2r_+\cdot(1-r_+) + r_+ - 1 = 0$$
$$\Leftrightarrow 2r_+ - 2r_+^2 + r_+ - 1 = 0$$
$$\Leftrightarrow -2r_+^2 + 3r_+ - 1 = 0$$
$$\Delta = 9 - 4\cdot(-2)\cdot(-1) = 1$$
$$r_{+1,2} = \frac{-3 \pm 1}{-4} \; \Big< \; \begin{array}{l} r_{1+} = 1 \text{ rejected } (r_+ < 1) \\ \boxed{r_{2+} = \frac{1}{2}} \end{array}$$

B) $1 - r_+ < 0 \Rightarrow \boxed{r_+ > 1}$ then $|1-r_+| = r_+ - 1$

So $$\frac{\partial L}{\partial r_+} = 0 \Leftrightarrow 2r_+(r_+ - 1) + r_+ - 1 = 0$$
$$\Leftrightarrow 2r_+^2 - 2r_+ + r_+ - 1 = 0$$
$$\Leftrightarrow 2r_+^2 - r_+ - 1 = 0$$

$$\Delta = 1 - 4\cdot(2)\cdot(-1) = 9$$
$$r_{+1,2} = \frac{1 \pm 3}{4} \; \Big< \; \begin{array}{l} r_{1+} = 1 \\ r_{2+} = -\frac{1}{2} \end{array}$$ both solutions are rejected as $r_+ > 1$

So, for $\boxed{\lambda = 1}$ the only representative of class + is $\boxed{r_+ = \frac{1}{2}}$

$\boxed{\lambda=2}$ $L(r_+) = r_+^2 + 1 + 2|1-r_+|$

$$\frac{\partial L}{\partial r_+} = 2r_+ + \frac{2\cdot(1-r_+)\cdot(-1)}{|1-r_+|} = 2r_+|1-r_+| + 2(r_+-1)$$

A) $1-r_+>0 \Rightarrow \boxed{r_+<1}$

$$\frac{\partial L}{\partial r_+} = 0 \Leftrightarrow 2r_+(1-r_+) + 2r_+ - 2 = 0$$
$$\Leftrightarrow 2r_+ - 2r_+^2 + 2r_+ - 2 = 0$$
$$\Leftrightarrow -2r_+^2 + 4r_+ - 2 = 0$$
$$\Leftrightarrow r_+^2 - 2r_+ + 1 = 0 \Leftrightarrow (r_+-1)^2 = 0 \Leftrightarrow r_+ - 1 = 0$$
$$\Leftrightarrow r_+ = 1$$

rejected $(r_+<1)$

B) $1-r_+<0 \Rightarrow \boxed{r_+>1}$

$$\frac{\partial L}{\partial r_+} = 0 \Leftrightarrow 2r_+(r_+-1) + 2r_+ - 2 = 0$$
$$\Leftrightarrow 2r_+^2 - 2r_+ + 2r_+ - 2 = 0$$
$$\Leftrightarrow 2r_+^2 - 2 = 0 \Leftrightarrow r_+^2 = 1 \Rightarrow r_+ = \pm 1$$ both solutions are rejected $(r_+>1)$

So, for $\lambda=2$, we can't accept any representation [illegible]

$\boxed{\lambda=3}$ $L(r_+) = r_+^2 + 1 + 3|1-r_+|$

$$\frac{\partial L}{\partial r_+} = 2r_+ + \frac{3(1-r_+)\cdot(-1)}{|1-r_+|} = 2r_+|1-r_+| + 3(r_+-1)$$

A) $1-r_+>0 \Rightarrow \boxed{r_+<1}$ $\frac{\partial L}{\partial r_+} = 0 \Leftrightarrow 2r_+(1-r_+) + 3r_+ - 3 = 0$
$$\Leftrightarrow 2r_+ - 2r_+^2 + 3r_+ - 3 = 0$$
$$\Leftrightarrow -2r_+^2 + 5r_+ - 3 = 0$$

$\Delta = 25 - 4(-2)\cdot(-3) = 1$

$r_{+1,2} = \frac{-5\pm1}{-4}$ $<$ $r_{+1} = \frac{3}{2}$, $r_{+2} = 1$ both solutions are rejected $(r_+>1)$

B) $1-r_+<0 \Rightarrow \boxed{r_+>1}$ $\frac{\partial L}{\partial r_+} = 0 \Leftrightarrow 2r_+(r_+-1) + 3r_+ - 3 = 0$
$$\Leftrightarrow 2r_+^2 - 2r_+ + 3r_+ - 3 = 0$$
$$\Leftrightarrow 2r_+^2 + r_+ - 3 = 0$$

$\Delta = 1 - 4\cdot2(-3) = 25$

$r_{+1,2} = \frac{-1\pm5}{4}$ $<$ $r_{+1} = -3/2$, $r_{+2} = 1$ both solutions are rejected $(r_+>1)$

So for $\lambda=3$, we can't accept any representation for class +

Hence, $\boxed{\lambda = 0} \rightarrow \boxed{r_+ = 0}$

$\boxed{\lambda = 1} \rightarrow \boxed{r_+ = \frac{1}{2}}$

$\boxed{\lambda = 2}$ → we can't accept any representor due to constraints

$\boxed{\lambda = 3}$ → we can't accept any representor due to constraints

The minimum values at these points are:

$\lambda = 0 : r_+ = 0$

$\boxed{L(0) = 1}$

$\lambda = 1 : r_+ = \frac{1}{2}$

$\boxed{L\left(\frac{1}{2}\right)} = \left(\frac{1}{2}\right)^2 + 1 + \left|1 - \frac{1}{2}\right| = \frac{1}{4} + 1 + \frac{1}{2} = \boxed{\frac{7}{4}}$

It is something that we expect from the graphs that we plotted in the previous question (Q1a)

Q2 The regularizer in Equation 1 tries to control the overfitting. So, it has to do with the stability and it affects the generalization performance of the classifier. So its presence is necessary and especially in higher dimensions.

In case that λ gets larger and larger the performance which we are going to obtain would be poor and the typical value of the coefficients get smaller. Also, we see that λ controls the effective complexity of the model. To be more precise, in the limiting case that $\lambda \rightarrow +\infty$, then in the loss function the term $\lambda \| r_- - r_+ \|_1$ becomes the dominant one. So the value of the loss function is going to be very large (opposite result of the desired one). So, in order to achieve our goal and to minimize the loss function, the quantity inside the L1 norm $(r_- - r_+)$ shall goes to zero. Hence the two representors should be equal $(r_- = r_+)$. Hence our classifier is going to have a bad performance as the two representors are going to lie in the same point.

the degree of overfitting

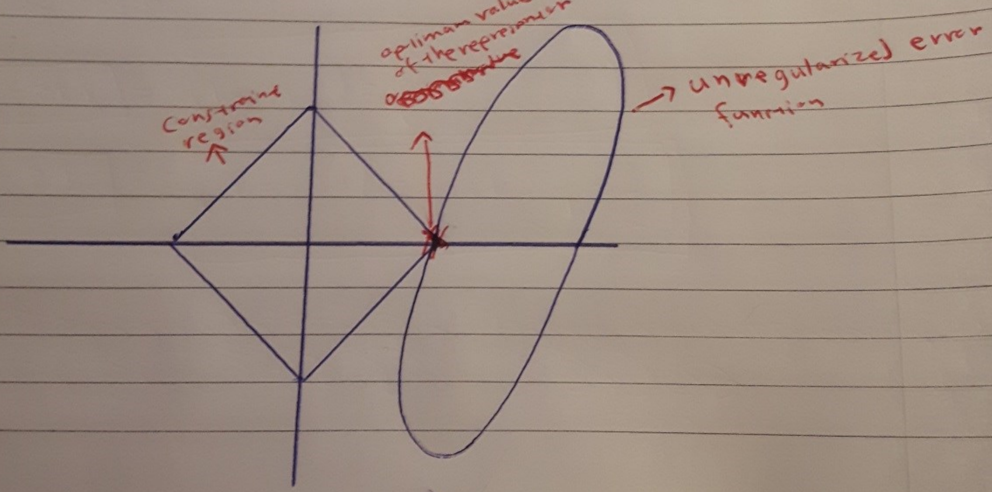
Q3 (a)

We have the loss function $L(r_-, r_+) = \sum_{i=1}^{N} \frac{1}{N_{y_i}} \|x_i - r_{y_i}\|^2 + \lambda \|r_- - r_+\|_1$

The first term of the loss function $\sum_{i=1}^{N} \frac{1}{N_{y_i}} \|x_i - r_{y_i}\|^2$ is like having a term in form of $\sum_{i=1}^{N} (f(x_i, w) - y_i)^2$ which is an ellipses.

The second term of the loss function $\lambda \|r_- - r_+\|_1$ is like having a term in form of $\lambda . \|w\|_1$ which is a Diamond (Lasso, as $p = 1$)

So when we are trying to find the 1-dimensional representors, the contour lines for the loss function L could be described as the concatenation of an ellipses and a "diamond" (Lasso)



Q3 b $+: x_1 = -1,\ x_2 = 1$

$-: x_3 = 3,\ x_4 = -1$

$\boxed{\lambda = 0}$

$\boxed{L(r_-, r_+)} = \sum_{i=1}^{4} \frac{1}{n_{y_i}} \|x_i - r_{y_i}\|^2$

$= \frac{1}{2}\|-1 - r_+\|^2 + \frac{1}{2}\|1 - r_+\|^2 + \frac{1}{2}\|3 - r_-\|^2 + \frac{1}{2}\|-1 - r_-\|^2$

$= \frac{1}{2}(-1-r_+)^2 + \frac{1}{2}(1-r_+)^2 + \frac{1}{2}(3-r_-)^2 + \frac{1}{2}(-1-r_-)^2$

$= r_+^2 + 1 + \frac{1}{2}(9 - 6r_- + r_-^2) + \frac{1}{2}(r_-^2 + 2r_- + 1)$

$= r_+^2 + 1 + \frac{9}{2} - 3r_- + \frac{r_-^2}{2} + \frac{r_-^2}{2} + r_- + \frac{1}{2}$

$= \boxed{r_+^2 + r_-^2 - 2r_- + 6}$

$\frac{\partial L(r_+, r_-)}{\partial r_+} = 0 \iff 2r_+ = 0 \implies \boxed{r_+ = 0}$

$\frac{\partial L(r_+, r_-)}{\partial r_-} = 0 \iff 2r_- - 2 = 0 \iff 2r_- = 2 \iff \boxed{r_- = 1}$

So $(r_-, r_+) = (1, 0)$ which is the expected solution

$\boxed{\lambda = \text{large enough}}$

$L(r_+, r_-) = r_+^2 + r_-^2 - 2r_- + 6 + \lambda\|r_- - r_+\|_1$
$= r_+^2 + r_-^2 - 2r_- + 6 + \lambda|r_- - r_+|$

$\frac{\partial L}{\partial r_+} = 0 \iff 2r_+ + \frac{\lambda(r_- - r_+)(-1)}{|r_- - r_+|} = 2r_+ - \frac{\lambda(r_- - r_+)}{|r_- - r_+|}$

Again we are going to distinguish 2 cases.

A) If $r_- - r_+ > 0 \iff \boxed{r_- > r_+} \implies |r_- - r_+| = r_- - r_+$

Hence, $\frac{\partial L}{\partial r_+} = 0 \iff 2r_+ \frac{-\lambda(r_- - r_+)}{r_- - r_+} = 0 \iff 2r_+ - \lambda = 0$

$\implies \boxed{r_+ = \frac{\lambda}{2}}$

B). If $r_- - r_+ < 0 \iff \boxed{r_+ > r_-} \Rightarrow |r_- - r_+| = r_+ - r_-$

So $\frac{\partial L}{\partial r_+} = 0 \iff 2r_+ - \frac{\lambda(r_- - r_+)}{r_+ - r_-} = 0 \iff 2r_+ - \frac{\lambda(r_- - r_+)}{-(r_- - r_+)} = 0$

$\iff 2r_+ + \lambda = 0 \iff \boxed{r_+ = -\frac{\lambda}{2}}$

After that we are going to take partial derivative with respect to r_-

$\frac{\partial L}{\partial r_-} = 0 \iff 2r_- - 2 + \frac{\lambda(r_- - r_+)}{|r_- - r_+|}$

A). If $r_- - r_+ > 0 \iff \boxed{r_- > r_+} \Rightarrow |r_- - r_+| = r_- - r_+$

So, $\frac{\partial L}{\partial r_-} = 0 \iff 2r_- - 2 + \frac{\lambda(r_- - r_+)}{(r_- - r_+)} = 0 \iff 2r_- - 2 + \lambda = 0$

$\iff 2r_- = 2 - \lambda$

$\iff \boxed{r_- = 1 - \frac{\lambda}{2}}$

B) If $r_- - r_+ < 0 \iff \boxed{r_+ > r_-} \Rightarrow |r_- - r_+| = r_+ - r_-$

So $\frac{\partial L}{\partial r_-} = 0 \iff 2r_- - 2 + \frac{\lambda(r_- - r_+)}{r_+ - r_-} = 0$

$\iff 2r_- - 2 + \frac{\lambda(r_- - r_+)}{-(r_- - r_+)} = 0 \iff 2r_- - 2 - \lambda = 0$

$\iff 2r_- = \lambda + 2$

$\iff \boxed{r_- = 1 + \frac{\lambda}{2}}$

Hence, we have that:

- $\boxed{r_- > r_+} \Rightarrow \boxed{(r_-, r_+) = \left(1 - \frac{\lambda}{2}, \frac{\lambda}{2}\right)}$
- $\boxed{r_+ > r_-} \Rightarrow \boxed{(r_-, r_+) = \left(1 + \frac{\lambda}{2}, -\frac{\lambda}{2}\right)}$

04 a

We are going to implement an optimizer ~~of~~ for the NrC. For this reason we are going to employ the Gradient descent method.

Step 1 Start with $\vartheta = (\vartheta_0, \dots, \vartheta_{64}) = 0$

Step 2 Keep changing $\vartheta_0, \dots, \vartheta_{64}$ until the time that we got the minimum value of $J(\vartheta_0, \dots, \vartheta_{64})$, where J is our loss function (1)

Update

Step 3 $temp_0 = \vartheta_0 - a \frac{\partial}{\partial \vartheta_0} J(\vartheta_0, \dots, \vartheta_{63})$

$\vdots$

$temp_j = \vartheta_j - a \frac{\partial}{\partial \vartheta_j} J(\vartheta_0, \dots, \vartheta_{63}) \quad j = 1, \dots, 63$

Step 4 $\vartheta_0 = temp_0$

Step 5 $\vartheta_j = temp_j$

So we begin to minimize the function $L(r_+, r_-)$ with respect to r_+ and r_- respectively.

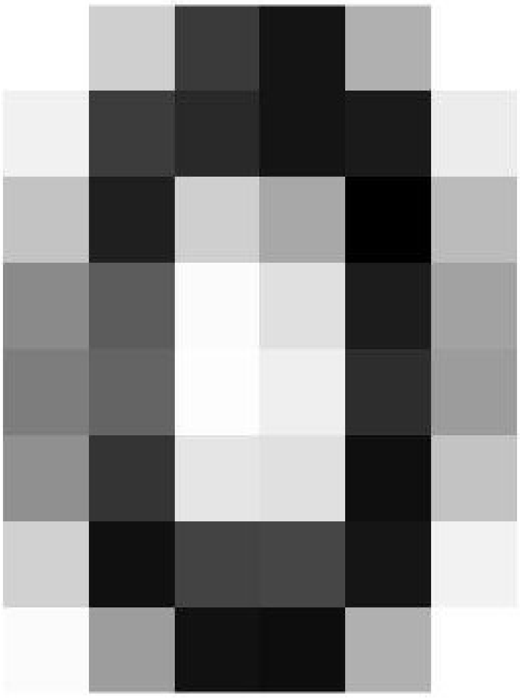
$$\partial L(r_-, r_+) = 0$$

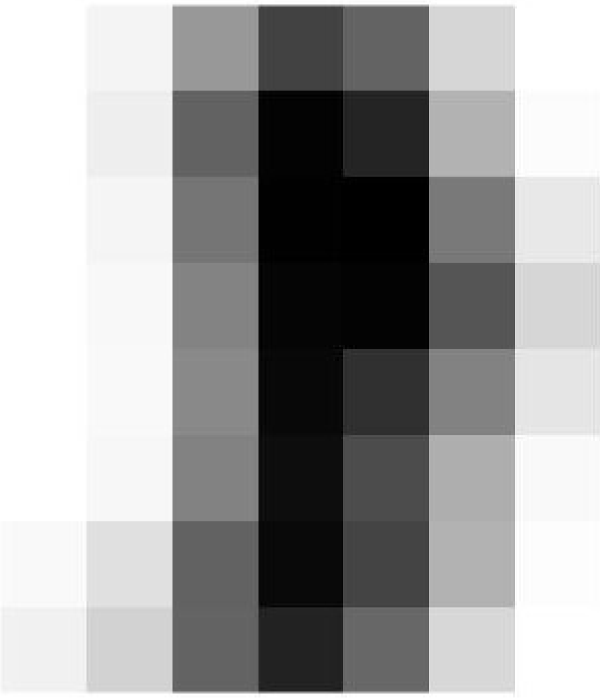
$$\frac{\partial L(r_+, r_+)}{\partial r_-} = 0 \qquad \frac{\partial L(r_-, r_+)}{\partial r_+} = 0$$

$$\frac{\partial L(r_+, r_+)}{\partial r_{-0}} = 0 \quad \dots \quad \frac{\partial L(r_+, r_-)}{\partial r_{-63}} = 0 \qquad \frac{\partial L(r_-, r_+)}{\partial r_{+0}} = 0 \quad \dots \quad \frac{\partial L(r_-, r_+)}{\partial r_{+63}} = 0$$

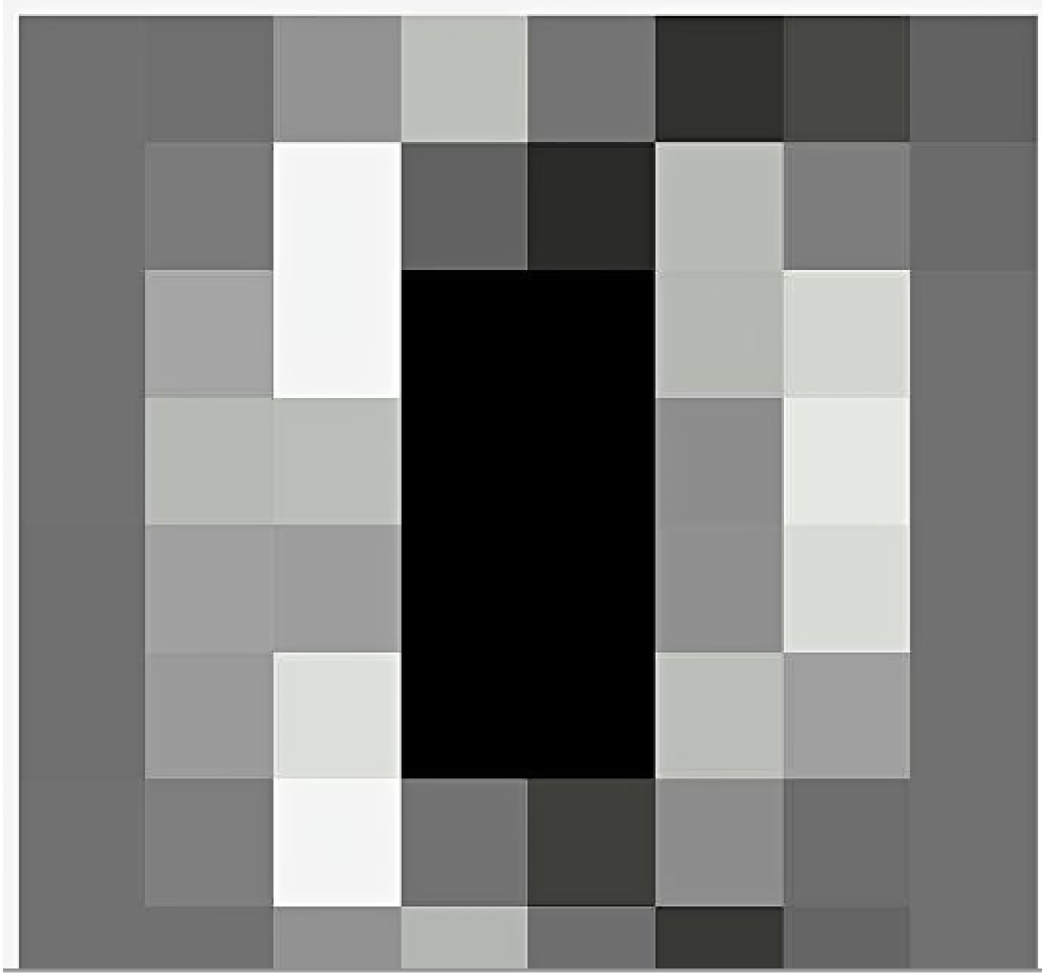
And by using an optimization toolbox of Matlab and function fminunc we are able to find the representors r_+ and r_- of each of 2 classes which are 64-dim arrays that minimize the loss function.

04b) In this question we plot the two representors of each class. In the two first figures below, the two representors of each class can be depicted by employing the procedure which we told in the previous question (4a) and minimizing the loss function through taking partial derivatives with respect to two representors and running it in MATLAB for λ=0.

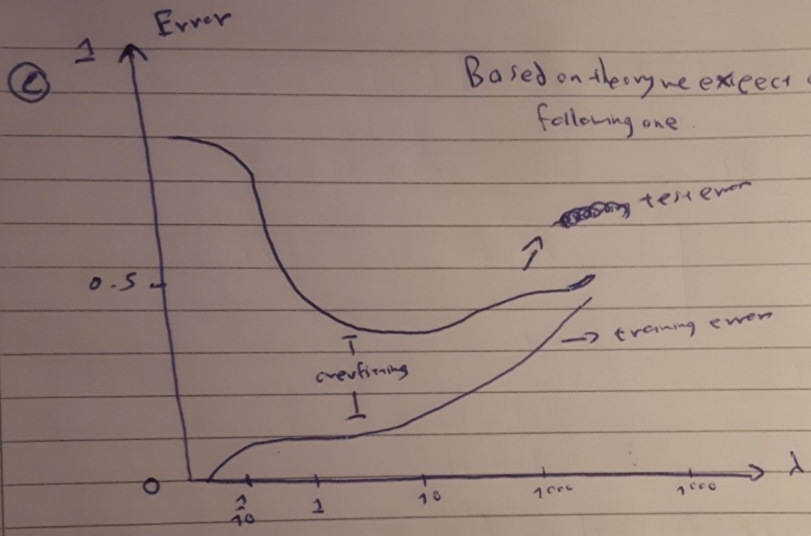




After that, we follow the same procedure for large values of λ. We notice that for λ greater than10^6 the solution does not change. Also it can be depicted that the two representors have come close enough due to the reason that we explain in question 2. So, it is obvious in the next figure that the 1(representor of one class) is inside the 0(representor of the other class).



Based on theory we expect a graph like the following one.



We see that λ controls the effective complexity of the model and determines the degree of overfitting