Question 1

a. A quadratic minimization problem that is equivalent to the given problem:

Minimize $\lambda\left||w|\right|^2 + \sum_{i=1}^{m} \xi_i^{\ 2}$

s.t. $\forall\, 1 \le i \le m,\ y_i\langle w, x_i\rangle \ge 1 - \xi_i\ \ \boldsymbol{and}\ \ \xi_i \ge 0$

$$if\ y_i\langle w, x_i\rangle \ge 1, \qquad then\ \xi_i^2 = 0^2 = \left[l^h\left(w, (x_i, y_i)\right)\right]^2$$

$$if\ y_i\langle w, x_i\rangle < 1, \qquad then\ \xi_i^2 = (1 - y_i\langle w, x_i\rangle)^2 = \left[l^h\left(w, (x_i, y_i)\right)\right]^2$$

b. z is of the form: $(w_1, \dots, w_d, \xi_1, \dots, \xi_m) \in \mathbb{R}^{d+m}$. So we have to set:

H is a block matrix of dimension $(d + m\ X\ d + m)$: $\forall\, 1 \le i \le d:\ H_{i,i} = 2\lambda$; $\forall\, 1 \le i \le m: H_{d+i,d+i} = 2$; and all the rest are $0s$.

$u \in \mathbb{R}^{d+m}$ is of the form: $\left(\overbrace{0, \dots, 0}^{d+m\ times}\right)$

A is also a block matrix of dimension $(2m\ X\ d + m)$:

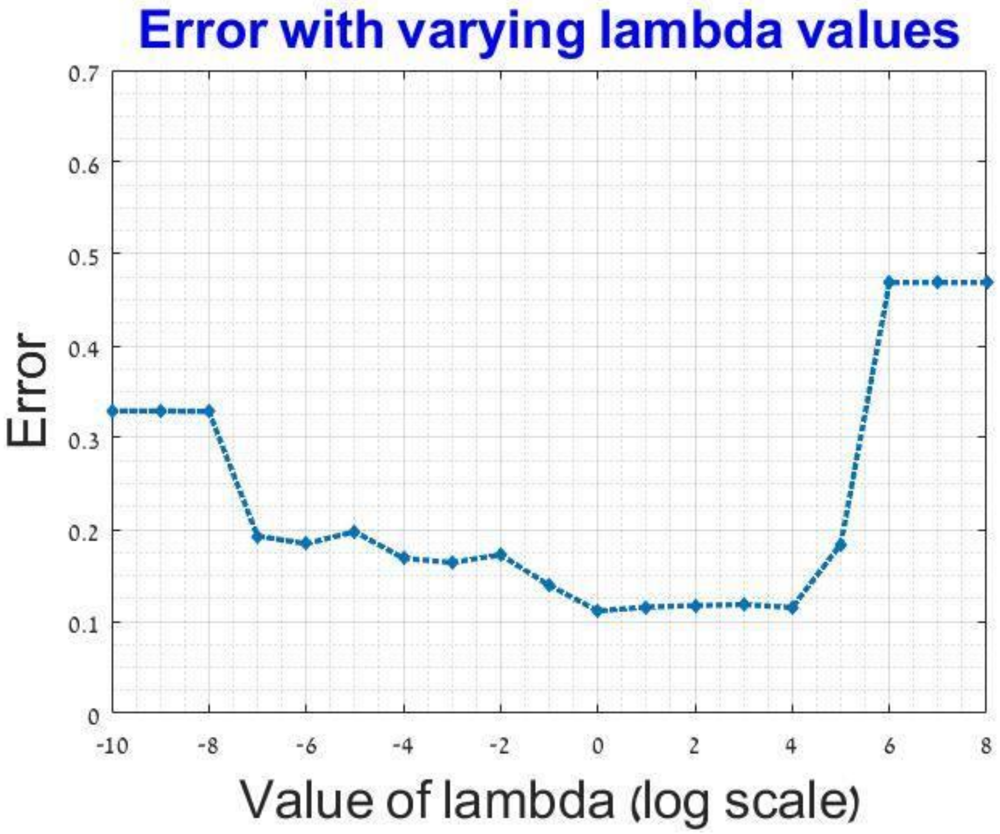
$$\forall\, 1 \le i \le m:\ row\ A_i = \left(y_i x_i(1), \dots, y_i x_i(d),\ \overbrace{0, \dots, 0}^{i-1\ times}\ , \mathbf{1}, 0, \dots 0\right)$$

$$and\ row\ A_{m+i} = \left(\overbrace{0, \dots, 0}^{d+i-1\ times}\ , \mathbf{1}, 0, \dots 0\right)$$

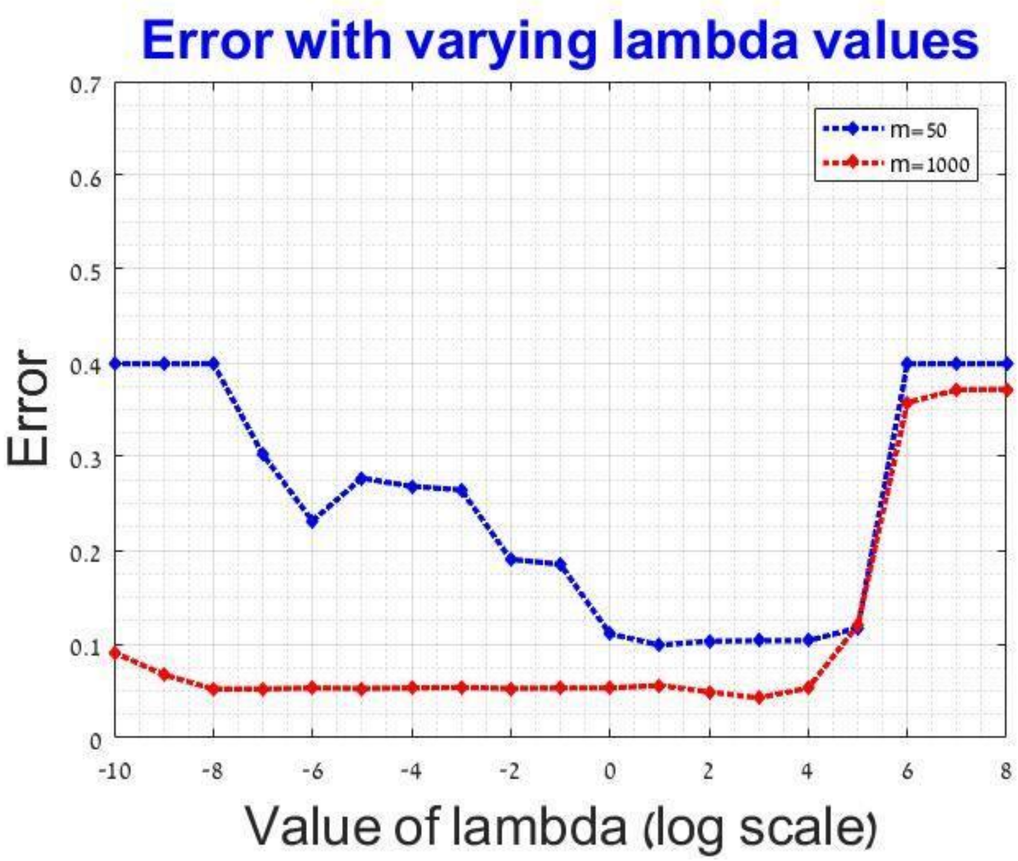
And $v \in \mathbb{R}^{2m} = \left(\overbrace{1, \dots, 1}^{m\ times}, \overbrace{0, \dots, 0}^{m\ times}\right)$

Question 3

a.



b.



c. Similarities: In both lines we see a relatively low error for $10^0 \leq \lambda \leq 10^5$.

Differences: In the blue line (sample size = 50), the error is high for both very small and very large values of lambda. In the red line (sample size = 1000), the error is low for small values of lambda, and is significantly higher for large values of lambda - $\lambda > 10^5$.

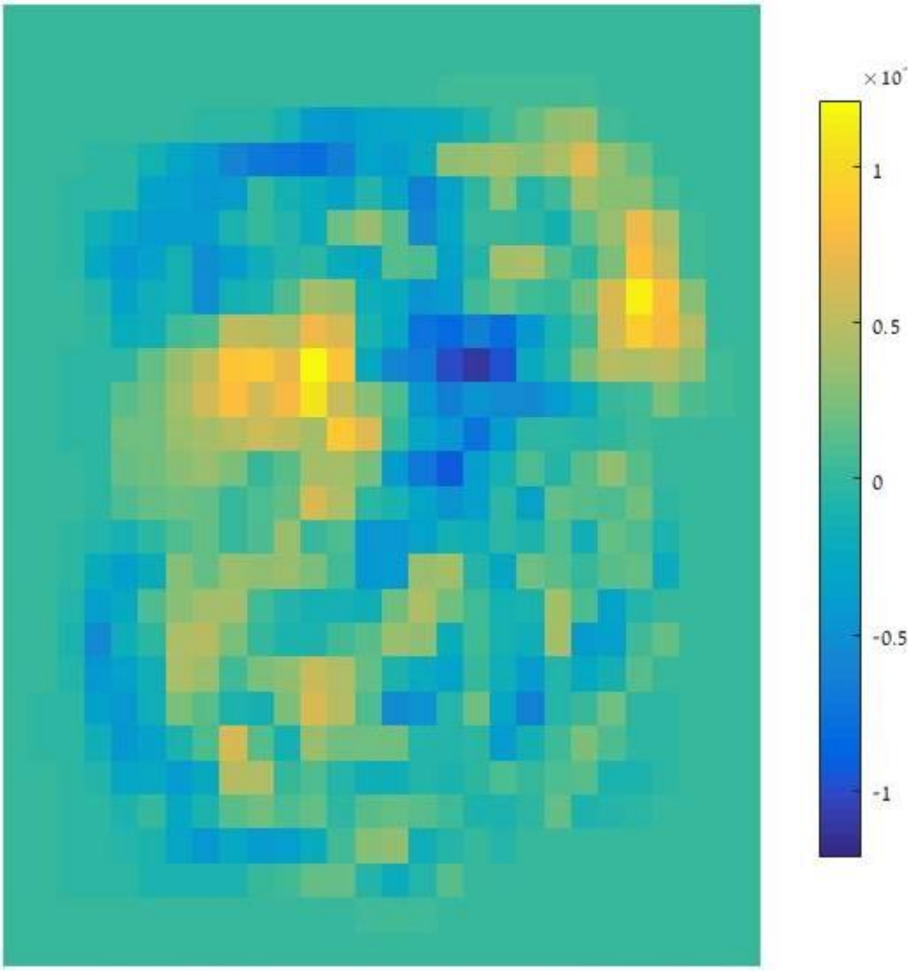
We explain our findings using what we learned in class regarding the trade-off in the value of lambda:

High error for small lambda with $m = 50$ can be explained by the fact that for small values of lambda, we "pay" more for small sample size, meaning high estimation error.

Low error for small lambda with $m = 1000$ reflects the same aspect of the trade-off: small lambda allows a big norm of w, which means small margin. So the hinge loss is with respect to a small soft-margin, therefore it is lower.

High error for large lambda in both sample sizes is explained by the fact that for large values of lambda, we get high approximation error, since we pay more for the norm of w - the minimization looks for w with smaller norm, which means large margin, so the hinge loss is higher. This does not depend on the sample size.

d.



e. Recall that the linear separator w is a vector of dimension $784\ X\ 1$, actually represents an image of $28\ X\ 28$ pixels. It is meant to separate images of the digit 3 from images of the digit 5 - every image of ‘3’ (label -1) should be classified on the “negative side” of w, and every image of ‘5’ (label 1) should be classified on the “positive side” of w. Since we marked negative values in blue and positive values in orange, we expect pixels that get high values in most training examples of the digit 3, to be blue in the image of w, and pixels that get high values in most training examples of the digit 5, to be orange in the image of w.

We notice that this is actually what happens in the image - the background consists of green pixels which represent the value 0. This is because most pixels in the background of both ‘3’ and ‘5’ are white (value 0). At the center of the image we can see that pixels that get high values in ‘3’ are painted in dark blue, pixels that get high values in ‘5’ are painted in bright orange, and pixels that get high values in both ‘3’ and ‘5’ are painted in green, which represents 0.

Question 4

a. Prove that for any i ≤ d, $|w^{t+1}(i)| \leq t$:

First we notice that by the definitions of x_i, y_i, for any i:

$$y_i x_i = (-1)^{i+1} * \left(\overbrace{(-1)^i, \dots, (-1)^i}^{i-1\ times}, (-1)^{i+1}, \overbrace{0, \dots, 0}^{d-i\ times}\right).$$

Now we prove the claim by induction on t, the Perceptron update $w^{t+1} \leftarrow w^t + y_i x_i$.

- Base case - $t = 1$: $|w^{t+1}(i)| = |w^2(i)| \underbrace{=}_{Perceptron\ update} |(w^1 + y_i x_i)(i)|$

$$w^1 + y_i x_i = (0, \dots, 0) + (-1)^{i+1} * \left(\overbrace{(-1)^i, \dots, (-1)^i}^{i-1\ times}, (-1)^{i+1}, \overbrace{0, \dots, 0}^{d-i\ times}\right)$$

$$= (0, \dots, 0) + \left(\overbrace{(-1)^{2i+1}, \dots, (-1)^{2i+1}}^{i-1\ times}, (-1)^{2i+2}, \overbrace{0, \dots, 0}^{d-i\ times}\right)$$

$\left|(-1)^{2i+1}\right| = |-1| = 1 \leq 1$

$\left|(-1)^{2i+2}\right| = |1| = 1 \leq 1$

$|0| = 0 \leq 1$

$\Rightarrow\ \forall i \leq d,\ |w^2(i)| \leq 1 = \boldsymbol{t}$

- Inductive step: Assume the induction hypothesis holds for any $k < t$, and prove for t:

$$|w^{t+1}(i)| = |(w^t + y_i x_i)(i)| = |w^t(i) + y_i x_i(i)| \leq |w^t(i)| + |y_i x_i(i)| \underbrace{\leq}_{\substack{induction\\ hypothesis}} t - 1 + |y_i x_i(i)| =$$

$$t - 1 + \begin{cases} \left|(-1)^{2i+1}\right| = |-1| = 1 \\ \left|(-1)^{2i+2}\right| = |1| = 1 \\ |0| = 0 \end{cases} \leq t - 1 + 1 = \boldsymbol{t}$$

b. Prove that for every i, $|w^T(i)| \geq 2^{i-1}$: We prove the claim by induction on i.

- Base case - $i = 1$: then $2^{i-1} = 2^0 = 1$, so we prove that $|w^T(1)| \geq 1$.

Assume that $|w^T(1)| < 1$.

Since the separator defined by w^T labels all the examples correctly, $y_1 * \langle w^T, x_1 \rangle > 0$.

By definition, $y_1 = (-1)^2 = 1$, so $\langle w^T, x_1 \rangle > 0$. But

$\langle w^T, x_1 \rangle = \langle w^T, ((-1)^2, \overbrace{0, \dots, 0}^{d-1\ times})\rangle = \langle w^T, (1, \overbrace{0, \dots, 0}^{d-1\ times})\rangle = w^T(1)$. So $w^T(1) > 0$, hence $|w^T(1)| > 0$.

But $w^T(1)$ has to be a whole number (we start with 0 and add -1 or 1 at each iteration), and we got

$0 < |w^T(1)| < 1$ - Contradiction.

- Inductive step: Assume the induction hypothesis holds for any $k < i$, and prove for i:

Since the separator defined by w^T labels correctly all the examples in S, $y_i \langle w^T, x_i \rangle > 0\ \forall i$.

From that and from the definition of x_i we get:

$$y_i \langle w^T, x_i \rangle = (-1)^{i+1} * \left[\sum_{k=1}^{i-1} w^T(k) * (-1)^i + w^T(i) * (-1)^{i+1}\right] =$$

$$\sum_{k=1}^{i-1} w^T(k) * (-1)^{2i+1} + w^T(i) * (-1)^{2i+2} \ = \sum_{k=1}^{i-1} -w^T(k) + w^T(i)$$

So:

$y_1\langle w^T, x_1\rangle = w^T(1) > 0$

$y_2\langle w^T, x_2\rangle = -w^T(1) + w^T(2) > 0 \ \Rightarrow \ w^T(2) > w^T(1) > 0$

$y_3\langle w^T, x_3\rangle = -w^T(1) - w^T(2) + w^T(3) > 0 \ \Rightarrow \ w^T(3) > w^T(1) + w^T(2) > 0$

And so on… we get: $w^T(i) > 0 \ \forall i \ \Rightarrow \ w^T(i) = |w^T(i)| \ \forall i.$

Now:

$$-\sum_{k=1}^{i-1} w^T(k) + w^T(i) > 0 \ \Rightarrow \ w^T(i) > \sum_{k=1}^{i-1} w^T(k) = \sum_{k=1}^{i-1} |w^T(k)| \underbrace{\geq}_{\substack{induction \\ hypothesis}} \sum_{k=1}^{i-1} 2^{k-1} = 2^{i-1} - 1$$

$$\Rightarrow \ w^T(i) > 2^{i-1} - 1 \quad \underbrace{\Rightarrow}_{\substack{w^T(i) \ is \ a \\ whole \ number}} \quad w^T(i) \geq 2^{i-1}$$

Therefore $|\boldsymbol{w^T(i)}| \geq \boldsymbol{2^{i-1}}$.

c. From the claim we proved in (b), for $i = d$ we get $|w^T(d)| \geq 2^{d-1}$. From that, combined with the fact that we start with $w = (0, \dots, 0)$ and in each iteration every coordinate i of w is increased by at most 1, we conclude that the number of updates the Perceptron makes until it stops is at least 2^{d-1} - exponential in d.

Assume this is not true, i.e, the Perceptron stops after $T < 2^{d-1}$ updates.

So according to the claim we proved in (a): $|w^T(i)| \leq T - 1 < 2^{d-1} - 1$, in <u>Contradiction</u> to $|w^T(d)| \geq 2^{d-1}$.