1.

(a)(i)

dataset 1:

<table>
<tr><td rowspan="2"></td><td colspan="3">Model Selection</td><td colspan="2">Performance</td></tr>
<tr><td>Best Param $\log_2\lambda$</td><td>Mean of MSE</td><td>Std of MSE</td><td>MSE on train</td><td>MSE on test</td></tr>
<tr><td rowspan="3">Least square</td><td>-</td><td>-</td><td>-</td><td>8.835</td><td>445.984</td></tr>
<tr><td rowspan="2">$\boldsymbol{w}$</td><td colspan="5">[1.324 0.777 -1.203 -1.721 -2.025 2.345 1.66 -2.694 -2.722 2.791]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 19.262</td><td>$\boldsymbol{l_2(w)}$ = 6.448</td><td colspan="2">Spars= 0</td></tr>
<tr><td rowspan="3">LASSO</td><td>-1</td><td>833.264</td><td>1455.607</td><td>0.122</td><td>433.811</td></tr>
<tr><td rowspan="2">$\boldsymbol{w}$</td><td colspan="5">[0.   0.   0.   -2.756 -1.387 3.792 0.   -5.641 -0.249 3.082]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 16.907</td><td>$\boldsymbol{l_2(w)}$ = 8.079</td><td colspan="2">Spars= 4</td></tr>
<tr><td rowspan="3">Ridge</td><td>5</td><td>1367.335</td><td>2334.431</td><td>37.658</td><td>639.490</td></tr>
<tr><td rowspan="2">$\boldsymbol{w}$</td><td colspan="5">[0.525 -0.194 -1.415 -1.503 -1.492 1.882 1.659 -2.239 -2.262 1.184]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 14.354</td><td>$\boldsymbol{l_2(w)}$ = 4.962</td><td colspan="2">Spars= 0</td></tr>
</table>

dataset 2

<table>
<tr><td rowspan="2"></td><td colspan="3">Model Selection</td><td colspan="2">Performance</td></tr>
<tr><td>Best Param $\log_2\lambda$</td><td>Mean of MSE</td><td>Std of MSE</td><td>MSE on train</td><td>MSE on test</td></tr>
<tr><td rowspan="3">Least square</td><td>-</td><td>-</td><td>-</td><td>61.264</td><td>135.502</td></tr>
<tr><td rowspan="2">$\boldsymbol{w}$</td><td colspan="5">[0.357 2.554 0.394 -6.085 3.204 3.319 0.788 -5.356 -2.923 1.435]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 26.415</td><td>$\boldsymbol{l_2(w)}$ = 10.248</td><td colspan="2">Spars= 0</td></tr>
<tr><td rowspan="3">LASSO</td><td>-1</td><td>99.746</td><td>44.055</td><td>62.347</td><td>128.615</td></tr>
<tr><td rowspan="2">$\boldsymbol{w}$</td><td colspan="5">[0.   2.532 0.335 -5.243 2.407 3.274 0.833 -6.128 -2.143 1.274]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 24.169</td><td>$\boldsymbol{l_2(w)}$ = 9.746</td><td colspan="2">Spars= 1</td></tr>
<tr><td rowspan="3">Ridge</td><td>3.5</td><td>97.098</td><td>42.230</td><td>63.591</td><td>125.633</td></tr>
<tr><td rowspan="2">$\boldsymbol{w}$</td><td colspan="5">[0.295 2.486 0.475 -4.798 2.015 2.751 1.375 -4.399 -3.85   1.311]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 23.755</td><td>$\boldsymbol{l_2(w)}$ = 8.883</td><td colspan="2">Spars= 0</td></tr>
</table>

dataset 3

<table>
<tr><td rowspan="2"></td><td colspan="3">Model Selection</td><td colspan="2">Performance</td></tr>
<tr><td>Best Param $\log_2\lambda$</td><td>Mean of MSE</td><td>Std of MSE</td><td>MSE on train</td><td>MSE on test</td></tr>
<tr><td rowspan="3">Least square</td><td>-</td><td>-</td><td>-</td><td>95.660</td><td>111.983</td></tr>
<tr><td rowspan="2">w</td><td colspan="4">[1.395 1.74   0.469 -2.83 -0.02   4.703 -0.139 -8.404 0.501 0.913]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 21.114</td><td>$\boldsymbol{l_2(w)}$ = 10.346</td><td colspan="2">Spars= 0</td></tr>
<tr><td rowspan="3">LASSO</td><td>-1.5</td><td>99.008</td><td>10.517</td><td>95.886</td><td>113.209</td></tr>
<tr><td rowspan="2">w</td><td colspan="4">[1.033 1.72   0.429 -2.816 -0.02   4.543 0.   -7.899 -0.   0.881]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 19.341</td><td>$\boldsymbol{l_2(w)}$ = 9.795</td><td colspan="2">Spars= 2</td></tr>
<tr><td rowspan="3">Ridge</td><td>1.5</td><td>99.989</td><td>9.566</td><td>95.677</td><td>111.844</td></tr>
<tr><td rowspan="2">w</td><td colspan="4">[1.391 1.739 0.471 -2.821 -0.031 4.652 -0.089 -8.069 0.166 0.913]</td></tr>
<tr><td>$\boldsymbol{l_1(w)}$ = 20.342</td><td>$\boldsymbol{l_2(w)}$ = 10.037</td><td colspan="2">Spars= 0</td></tr>
</table>

(ii)

(1)
N = 5

l1 regularizer < no regularizer < l2 regularizer

N = 50

l2 regularizer < l1 regularizer < no regularizer

N = 500

l2 regularizer < no regularizer < l1 regularizer

(2)

Both regularizers lower the norm of w (except l1 regularizer in N = 50 case does not lower l2 norm). When data is small, regularizers lower norm very much. When data is big, it does not lower very much. Because as data size increases, the result w of no regularizer, l1 regularizer and l2 regularizer become closer. So, norm of w in each case also become closer.

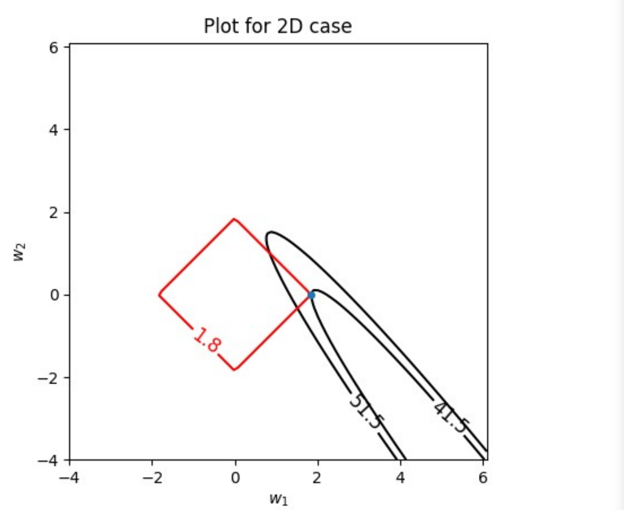
(3)
l1 regularizer will incur more sparsity because of regularization term.
Larger data my lead to less sparsity.
Larger lambda with the same data size will lead to more sparsity.

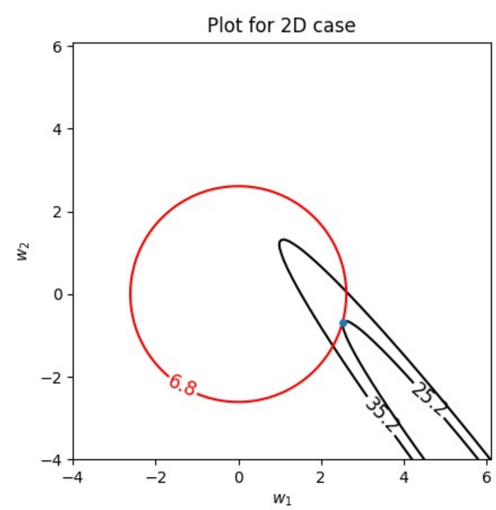
(b)

(i)

dataset 4:

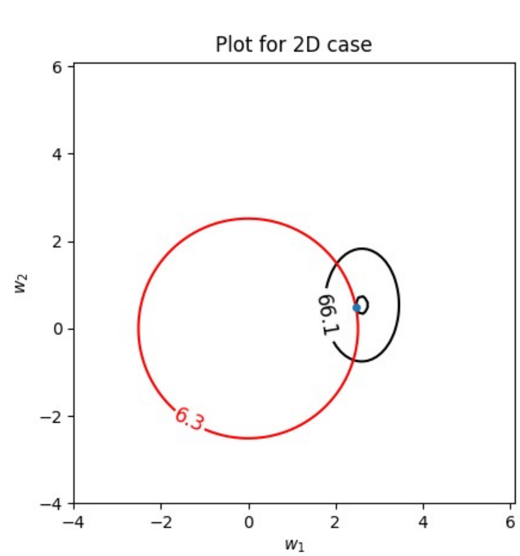
	Model Selection			Performance	
	Best Param $\log_2\lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	15.214	125.48
	w	[2.161 6.265 -5.463]			
		$\boldsymbol{l_1(w)}$ = 13.889	$\boldsymbol{l_2(w)}$ = 8.588	**Spars**= 0	
LASSO	2	63.114	69.578	41.462	109.641
	w	[0.216 1.849 0.]			
		$\boldsymbol{l_1(w)}$ = 2.065	$\boldsymbol{l_2(w)}$ = 1.861	**Spars**= 1	
Ridge	2	58.243	65.057	25.241	105.735
	w	[2.238 2.52 -0.684]			
		$\boldsymbol{l_1(w)}$ = 5.442	$\boldsymbol{l_2(w)}$ = 3.439	**Spars**= 0	





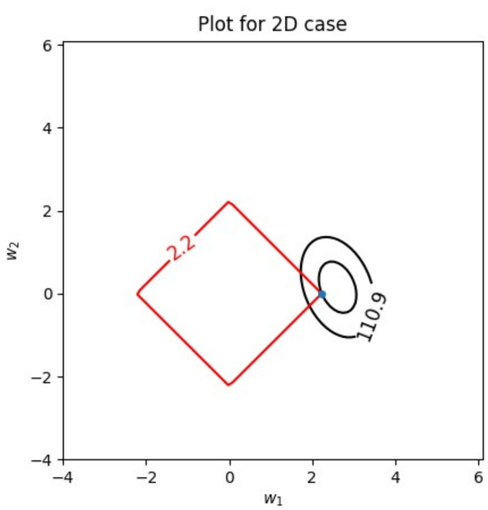
dataset 5:

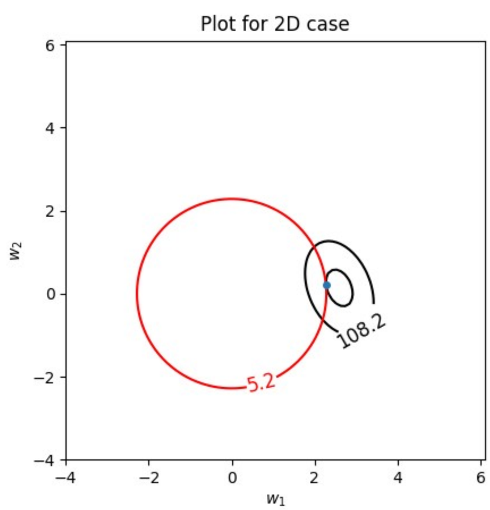
	Model Selection			Performance	
	Best Param $\log_2\lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	52.596	109.494
	w	[4.201 2.567 0.601]			
		$\boldsymbol{l_1(w)}$ = 7.369	$\boldsymbol{l_2(w)}$ = 4.959	**Spars**= 0	
LASSO	-1.5	78.036	21.202	52.757	107.062
	w	[3.837 2.549 0.53]			
		$\boldsymbol{l_1(w)}$ = 6.916	$\boldsymbol{l_2(w)}$ = 4.636	**Spars**= 0	
Ridge	3.5	70.982	37.002	56.085	101.059
	w	[2.394 2.47 0.476]			
		$\boldsymbol{l_1(w)}$ = 5.34	$\boldsymbol{l_2(w)}$ = 3.472	**Spars**= 0	



dataset 6:

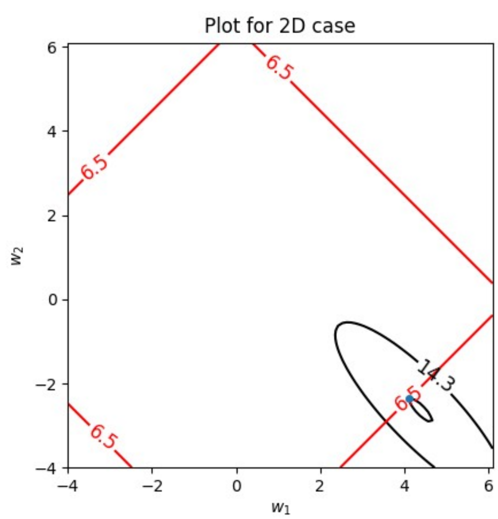
	Model Selection			Performance	
	Best Param $\log_2\lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	95.663	103.360
	w	[1.317 2.559 0.103]			
		$l_1(\boldsymbol{w})$ = 3.979	$l_2(\boldsymbol{w})$ = 2.879	**Spars**= 0	
LASSO	3	103.127	45.913	100.859	106.271
	w	[0. 2.229 0.]			
		$l_1(\boldsymbol{w})$ = 2.229	$l_2(\boldsymbol{w})$ = 2.229	**Spars**= 2	
Ridge	7	113.649	53.924	98.240	103.431
	w	[0.426 2.272 0.221]			
		$l_1(\boldsymbol{w})$ = 2.919	$l_2(\boldsymbol{w})$ = 2.322	**Spars**= 0	

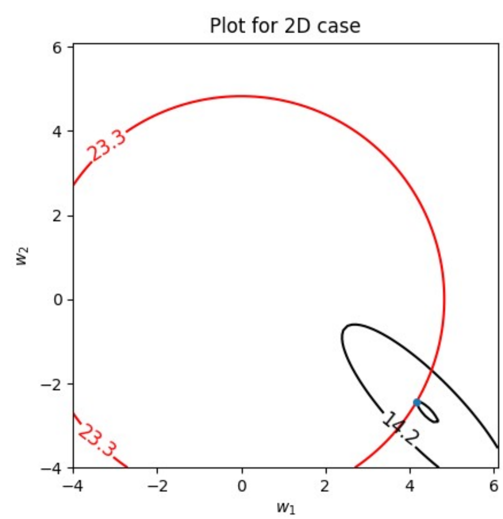




dataset 7:

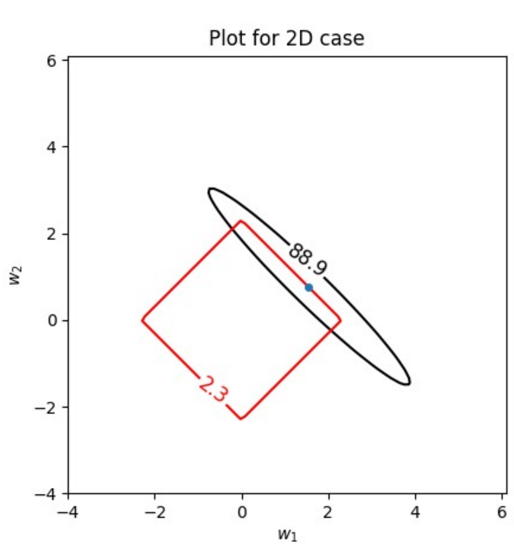
	Model Selection			Performance	
	Best Param $\log_2\lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	3.808	140.900
	w	[2.736 4.521 -2.814]			
		$\boldsymbol{l_1(w)}$ = 10.071	$\boldsymbol{l_2(w)}$ = 5.986	**Spars**= 0	
LASSO	-1.5	28.726	24.023	4.340	131.239
	w	[2.093 4.119 -2.358]			
		$\boldsymbol{l_1(w)}$ = 8.57	$\boldsymbol{l_2(w)}$ = 5.187	**Spars**= 0	
Ridge	-1.0	24.669	21.992	4.154	133.214
	w	[2.27 4.17 -2.432]			
		$\boldsymbol{l_1(w)}$ = 8.872	$\boldsymbol{l_2(w)}$ = 5.334	**Spars**= 0	

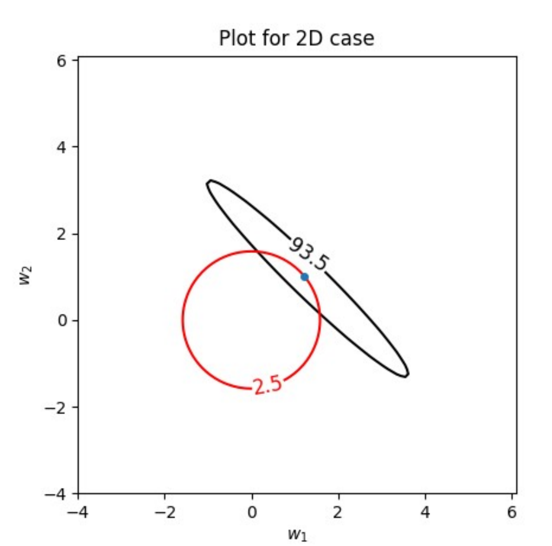




dataset 8:

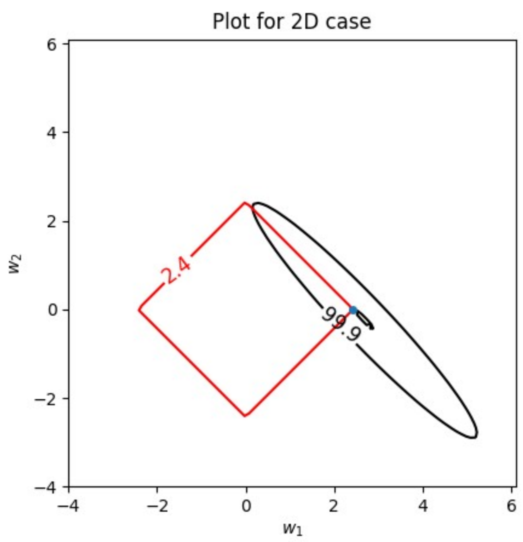
	Model Selection			Performance	
	Best Param $\log_2\lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	78.280	126.548
	w	[6.333 1.703 0.67]			
		$\boldsymbol{l_1(w)}$ = 8.706	$\boldsymbol{l_2(w)}$ = 6.592	**Spars**= 0	
LASSO	-0.5	145.548	110.751	78.914	121.406
	w	[5.498 1.551 0.757]			
		$\boldsymbol{l_1(w)}$ = 7.806	$\boldsymbol{l_2(w)}$ = 5.762	**Spars**= 0	
Ridge	3	135.957	107.437	83.454	115.667
	w	[3.922 1.221 1.009]			
		$\boldsymbol{l_1(w)}$ = 6.152	$\boldsymbol{l_2(w)}$ = 4.229	**Spars**= 0	

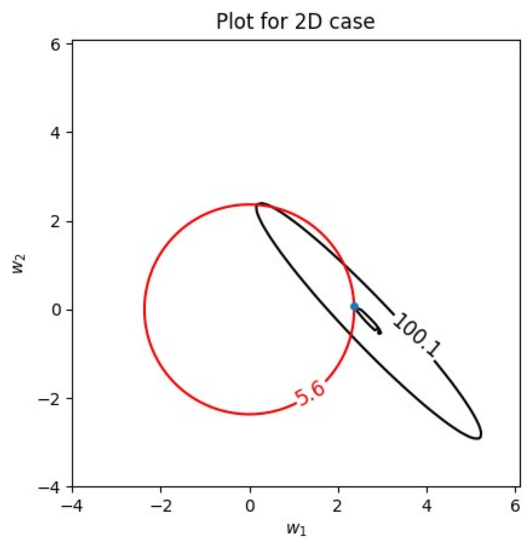




dataset 9:

	Model Selection			Performance	
	Best Param $\log_2\lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-	89.226	113.246
	w	[4.679 2.658 -0.183]			
		$\boldsymbol{l_1(w)}$ = 7.52	$\boldsymbol{l_2(w)}$ = 5.384	**Spars**= 0	
LASSO	-0.5	93.353	16.585	89.850	108.857
	w	[3.948 2.429 0.]			
		$\boldsymbol{l_1(w)}$ = 6.377	$\boldsymbol{l_2(w)}$ = 4.635	**Spars**= 1	
Ridge	3.5	95.800	16.149	90.129	108.053
	w	[3.806 2.367 0.071]			
		$\boldsymbol{l_1(w)}$ = 6.244	$\boldsymbol{l_2(w)}$ = 4.482	**Spars**= 0	

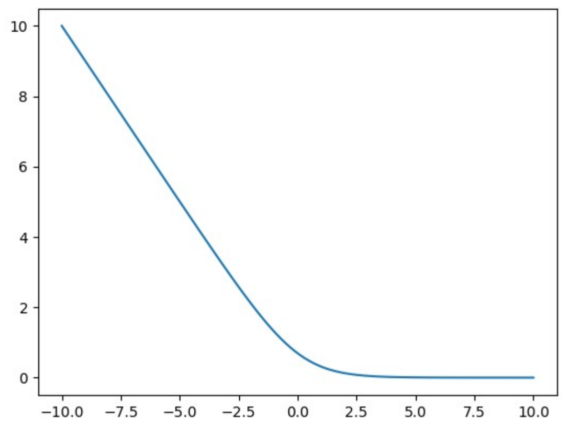


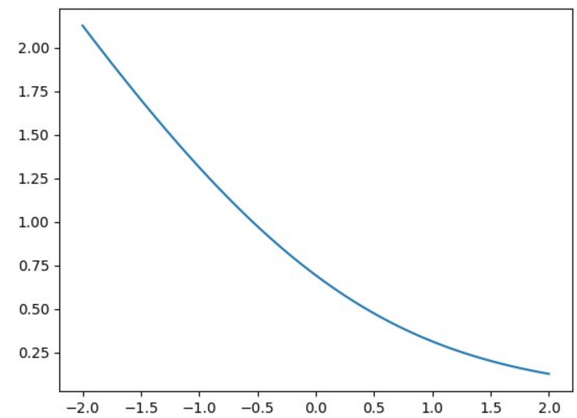


(iii)
(1) From the plots we can get predicted w and its coordinates. Each zero in its coordinates lead to one sparsity.
(2) The regularizer makes predicted result from center point of ellipse to the intersection point. If there is no regularizer, the result should be the center of ellipse in the plots.
(3) For dataset 5 and 8, l1 regularizer result is similar to result without regularizer. For other datasets, l1 regularizer takes a more obvious effects.

4.

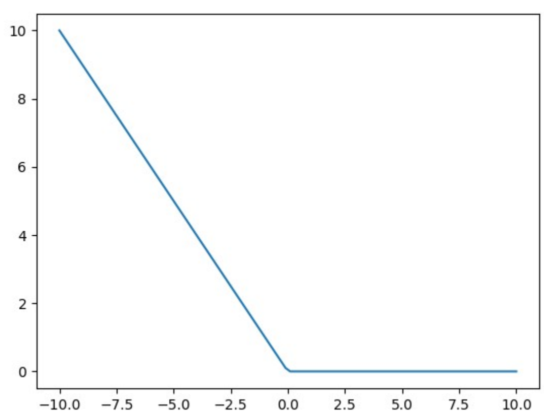
(a)

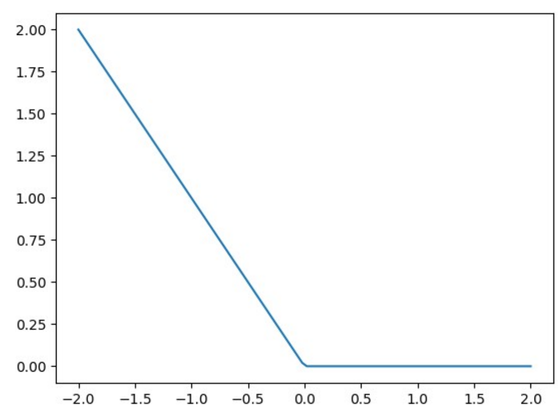




(b)

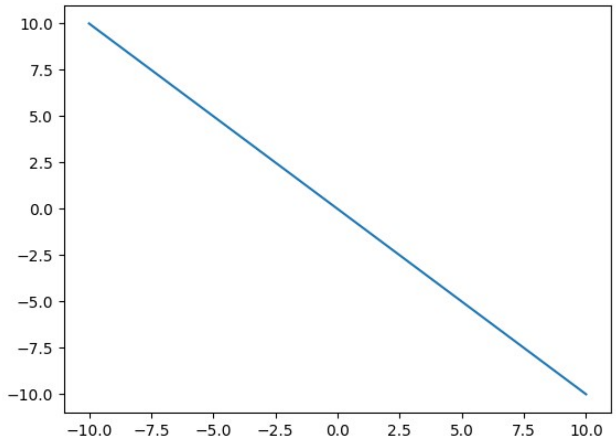
$E_i^{(p)} = -[s_i <= 0]\, s_i$

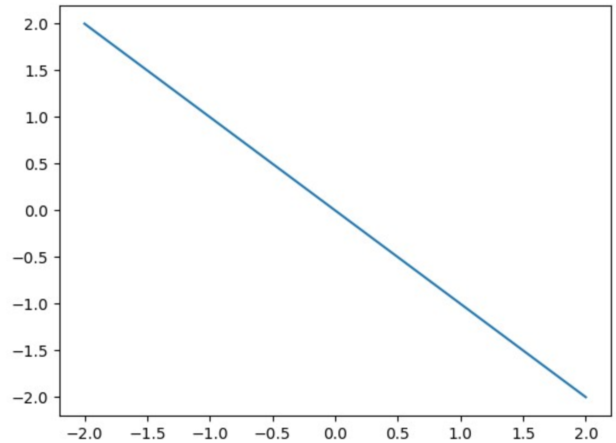




(c)

$E_i^{(MSE)} = (1/N)*(-2s)$





(d)

linear perceptron is more discriminative for those data points near decision boundary than logistic regression based on MLE.