Problem 2

After learning the regression part and different regularizations, Bob is interested in trying them out right away! He starts with the linear regression problem: given that the feature vector $\boldsymbol{x}$ and the observation y have a linear relationship $y = \boldsymbol{w}^T\boldsymbol{x} + w_0 + n$, estimate the weight vector $\boldsymbol{w} = [w_1, w_2, \ldots]$ and the bias w_0 from multiple data points. For simplicity in writing, we can augment the feature space and now parameters to be estimated can be written as $\boldsymbol{w} = [w_0, w_1, w_2, \ldots]$. Here n is the observation noise on the output labels y.

Bob starts to collect some samples to generate his dataset. He does the collection for several times and gets several datasets with different numbers of samples:

	number of training samples (N_{tr})	number of testing samples
Dataset1	10	1000
Dataset2	100	1000
Dataset3	1000	1000

Could you help him out on analysis on all the datasets above?

(a) Given that the dimension of features is 9 (before augmentation), estimate the $\boldsymbol{w}$ and try three regularization settings: [no regularization, l_1 regularization, l_2 regularization] and report the corresponding statistics. For each regularization setting to try, you need to search for a good regularization coefficient λ over the range $-10 \leq \log_2 \lambda \leq 10$ with step size of 0.5 for $\log_2 \lambda$, and use MSE (mean squared error) on the validation set to choose the best one. During the parameter search, you need to do 5-fold cross validation on each parameter value you try. **Tip:** after finding the best value of λ, use that value for one final training run using all N_{tr} training data points (nothing held out as a validation set), to get the weight vector and training MSE.

i. Fill all your numerical results into the following table. (Each dataset should have a different table. So for this question you'll have 3 tables.)

ii. Based on statistics on all datasets, answer the following questions:

1. Comparison of test MSE with no regularizer, l_1 regularizer, and l_2 regularizer for a given N_{tr} (your answer might also depend on N_{tr})
2. Does each regularizer lower the corresponding norm of $\boldsymbol{w}$? by very much? Please explain. Why are these answers different depending on N_{tr}?
3. Observe and explain the dependence of sparsity on regression method, and on different values of N_{tr} and λ.

	Model selection			Performance	
	Best param $\log_2 \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-		
	$\boldsymbol{w}$	(show your estimated w)			
		$l_1(w) =$	$l_2(w) =$	Spars=	
LASSO					
	$\boldsymbol{w}$	(show your estimated w)			
		$l_1(w) =$	$l_2(w) =$	Spars=	
Ridge					
	$\boldsymbol{w}$	(show your estimated w)			
		$l_1(w) =$	$l_2(w) =$	Spars=	

Caption for statistics in the table:

- Best param λ: the regularization coefficient you choose using cross validation.
- Mean of MSE: the averaged MSE of the 5-fold cross validation process for your chosen λ.
- Std of MSE: the standard deviation of MSE of the 5-fold cross validation process for your chosen λ.
- $l_1(w)$: l_1 norm of $\boldsymbol{w}$
- $l_2(w)$: l_2 norm of $\boldsymbol{w}$
- Spars: Sparsity, i.e., the number of zeros in the augmented weight vector

(b) Bob learned that l_1 regularization could lead to more sparsity, and he really wants to visualize this. So he collects another bunch of datasets for 2-dimensional (before augmentation) features:

	number of training samples (N_{tr})	number of testing samples
Dataset4	10	1000
Dataset5	30	1000
Dataset6	100	1000
Dataset7	10	1000
Dataset8	30	1000
Dataset9	100	1000

He tries them out and find that the last three datasets (7,8,9) are “special cases” where the l_1 norm might not provide the intended result.

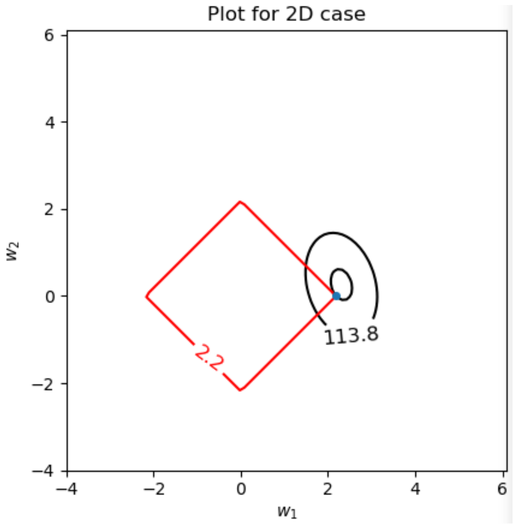
i. Repeat (a)(i) for all new datasets. (You’ll have 6 tables)

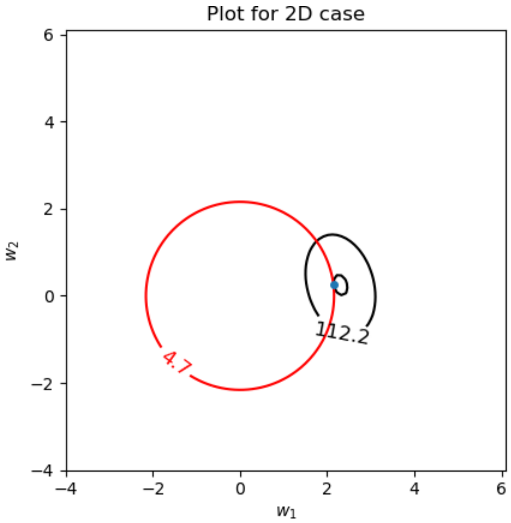
ii. For each dataset, draw the following plot in the 2D space w_2 vs. w_1 with $w_0 = \text{your estimated } w_0$: (1) draw the curve of ‘MSE = training_MSE of your estimated $\boldsymbol{w}$ and ‘MSE=10+training_MSE of your estimated $\boldsymbol{w}$; (2) draw the curve for $\|w\|_{l1} = \text{the } l_1 \text{ norm of your estimated } \boldsymbol{w}$. Repeat this plot drawing for ridge regression results, except for (2) draw the curve for $\|w\|_{l2} = \text{the } l_2 \text{ norm of your estimated } \boldsymbol{w}$. (therefore you have 2 plots for each dataset. An example is shown below.)

iii. Based on the statistics and plots, answer the following questions:

1. Observe and explain how the plots relate to sparsity.

2. Can you explain how much effect the regularizer has, from looking at the plots (i.e., how different the regularized performance (MSE) is from the unregularized performance)
3. Observe and explain how Lasso has a different effect with the “special case” datasets than the other datasets





Hint: please refer to the example code file in the homework folder on how to generate such plots.