Part 5

Question 1:

Add an l2 regularisation term to your objective function from Part 1 or Part 3. Compare the solution of the unregularized problem to those obtained while solving the problem with

- *a small value for the regularisation parameter and*
- *a large value for the regularisation parameter*

ℓ_2-norm - Ridge

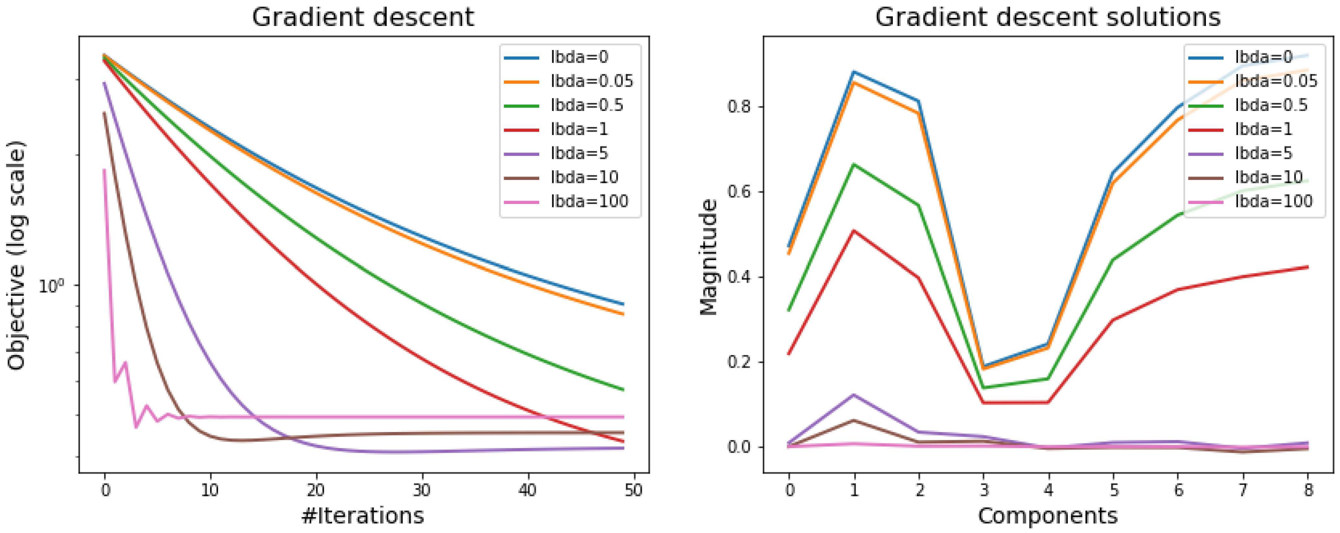
As in Part 1, the function to be minimised may be written as

$$\min_{\mathbf{x}\in\mathbb{R}^d} f_{\ell_2}(\mathbf{x}) = \frac{1}{2n}\|\mathbf{Ax}-\mathbf{y}\|^2 + \frac{\lambda}{2}\|x\|_2^2,$$

where $\lambda > 0$

We run several iterations for varying values of λ

```
Lambda  Norm of solution
0       2.1098
0.05    2.0342
0.5     1.4663
1       1.0233
5       0.1301
10      0.0654
100     0.0071
```



As discussed before, the addition of an ℓ_2-regularisation term enforces strong convexity and thus ensures that the problem has a unique global solution. We observe above the impact that λ has on the problem. As $\lambda \to \infty$, the regularisation term reduces the dependency of the solution to the data as it increasingly dominates the eqauation by its magnitude and thus also the variance with respect to the data.

We observe also the effect discussed in class that the introduction of ℓ_2 regularisation enforces a constraint on the ℓ_2 norm of the solution and thus drives the solution to have increasingly smaller values in termps of the ℓ_2-norm.

In other words, the addition of an ℓ_2-norm smoothes out the solution and leads to values of the solution vector that have increasingly smaller norms. We may once again observe the figure on the right above to see that the solution components shrink in a rather uniform (i.e. smooth) fashion.

Question 2 :

Add a l-1 regularisation term to your objective function from Part 1 or Part 3 and solve the resulting problem. Can you find a value of the regularisation parameter that yields a sparse solution? Does it provide a good value for the data-fitting term?

ℓ_1-norm - Lasso

We now add an ℓ_1 regularisation term to the data and consider the new problem

$$\min_{x \in \mathbb{R}^d} f_{\ell_1}(x) = \frac{1}{2n} ||Ax - y||^2 + \frac{\lambda}{2} ||x||_1,$$

The gradient of the regularisation term does not exist and so a typical solution is obtained by means of the ISTA-algorithm. The usual gradient descent step for the smooth part of the objective is given by:

$$g_k = x_k - \alpha_k \frac{1}{n} A^T (Ax - y).$$

For an iterate x_k and a stepsize $\alpha_k > 0$, each update to an iterate will be:

$$x_{k+1} = \mathbf{prox}_{\lambda \alpha_k} (x_k - \alpha_k g_k)$$

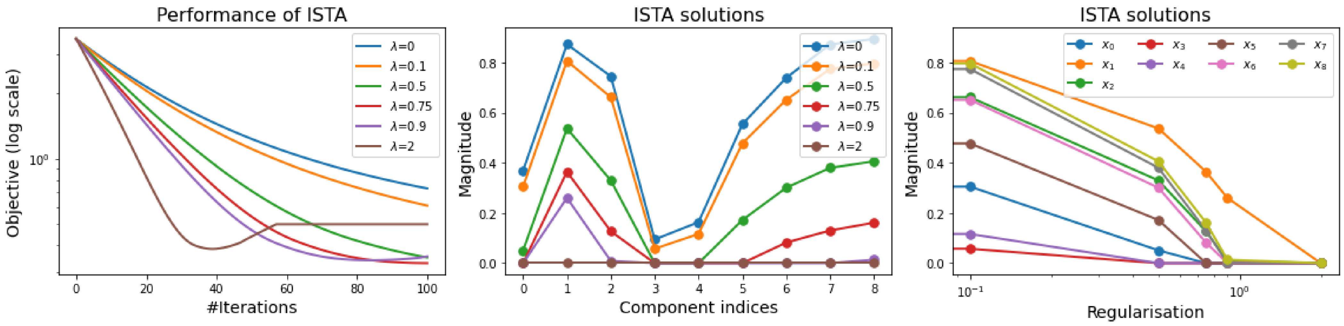
where the proximal form is computed by making use of the (coordinate-wise) soft-thresholding formula:

$$\forall i = 1, \ldots, d, \quad [x_{k+1}]_i \; = \; \begin{cases} [g_k]_i + \lambda \alpha_k & \text{if} [g_k]_i < -\lambda \alpha_k \\ [g_k]_i - \lambda \alpha_k & \text{if} []_i > \lambda \alpha_k \\ 0 & \text{otherwise.} \end{cases}$$

We consider the behavior of ISTA for several values of the regularisation parameter λ, namely:

$$\lambda \in \left\{ 0, \frac{1}{100}, \frac{1}{10}, \frac{1}{5}, \frac{1}{2} \right\}.$$

```
lbda     #non-zero
0:               9
0.1:             9
0.5:             7
0.75:            5
0.9:             3
2:               0
```

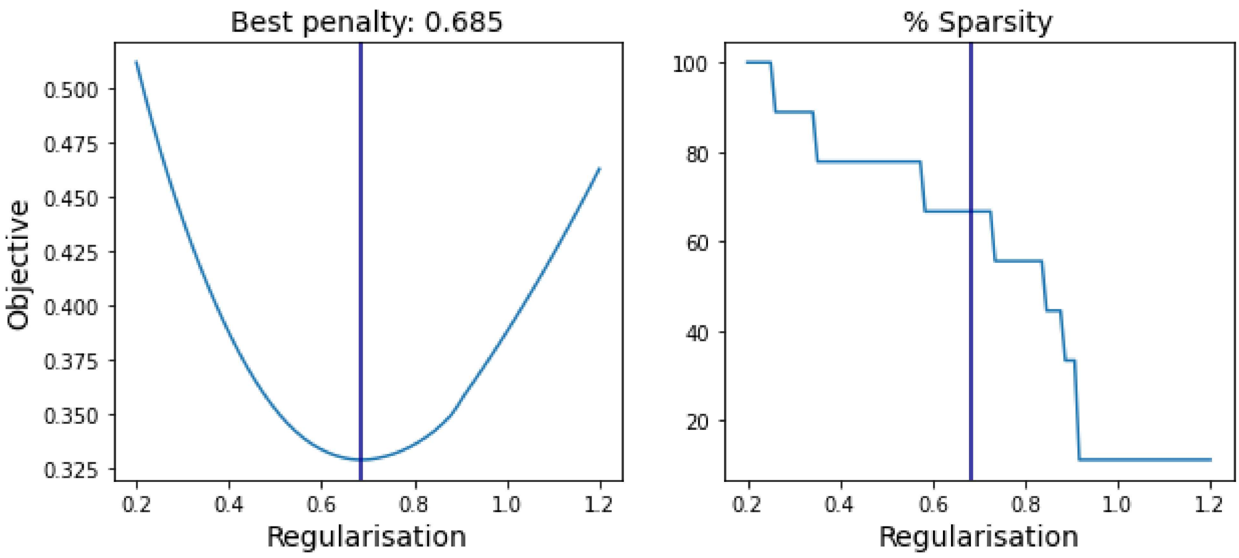


ℓ_1 regularisation does indeed lead to sparser solutions. Indeed, we observe that increasing regularisation parameter λ leads to a decrease in the number of non-zero components; and only the largest components remain. (See the middle graph above where component indices #3 and #4 are set to 0 'first', before indices such as #1 and #8 are)

For a regularisation parameter which is large enough, all components end up being set to 0 as the regularisation term dominates the expression to be minimised and hence having all components equal to zero is more beneficial than truly minimising the objective.

This is not always beneficial, as may be seen from the left-most figure above. Vectors which still have non-zero components provide better solutions than their all-zero counterparts.

Consider now a series of runs where the value of λ is varied on the interval $[0.2, 1.2]$. We find the best value on the data set and see that it is indeed attained with not all components being set to 0.

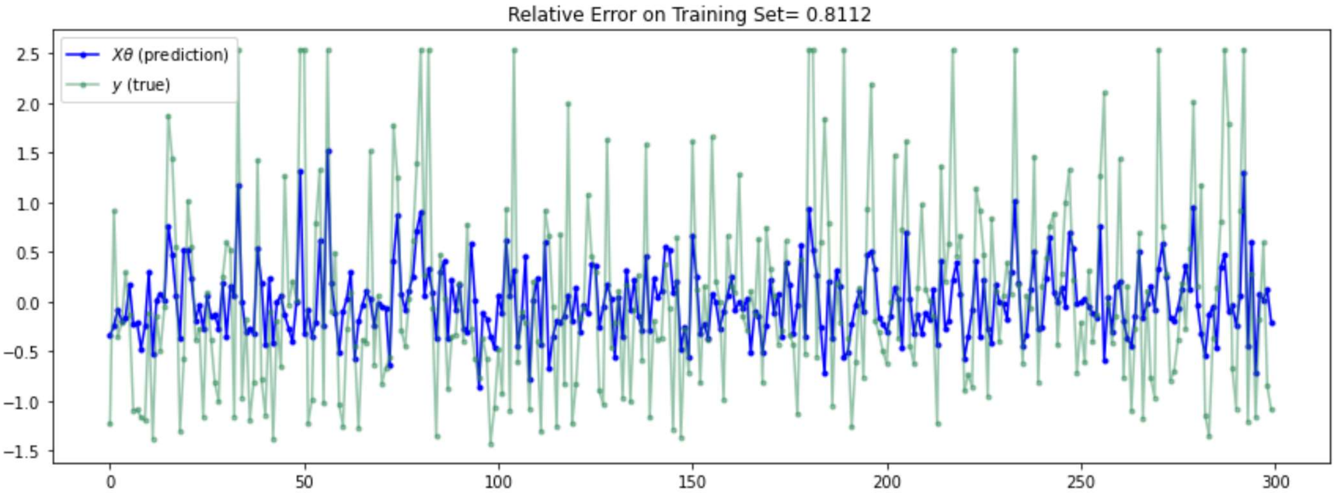


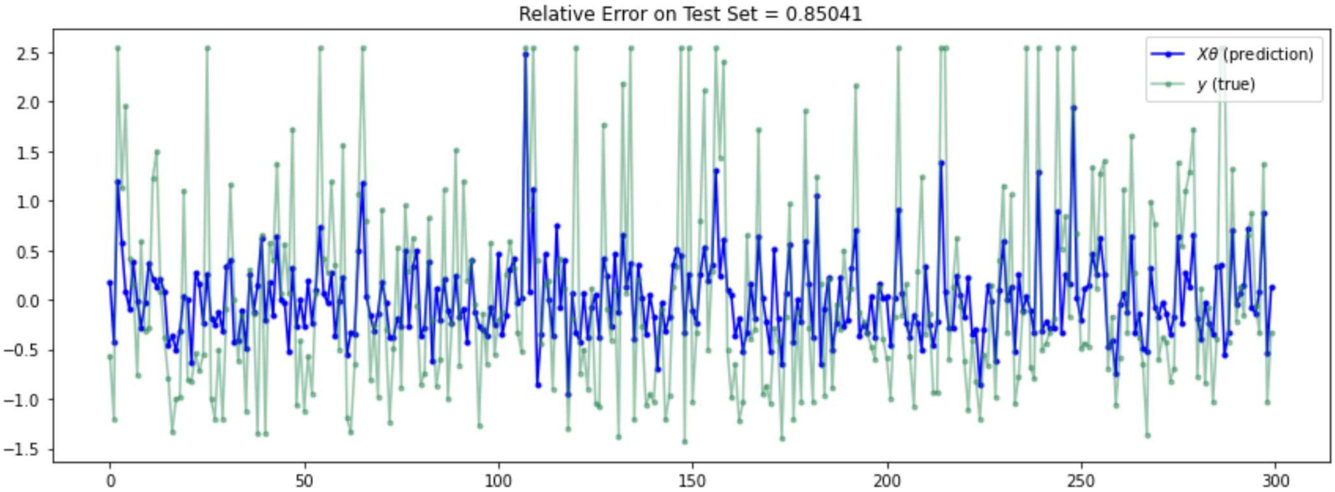
```
Number of non-zero components at optimal:
6
```

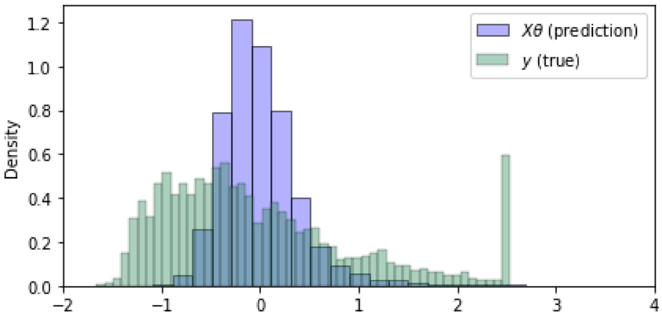
This implies that considering only a subset of the features will be sufficient. Specifically,

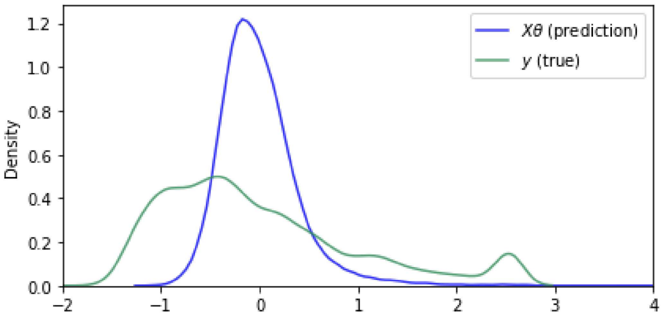
```
Features retained:
 ['HouseAge', 'AveRooms', 'AveOccup', 'Latitude', 'Longitude', 'Intercept']
Features removed
 ['MedInc', 'AveBedrms', 'Population']
```

Let us now use this to consider prediction performance


Relative Error on Training Set= 0.8112
$X\theta$ (prediction)
y (true)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
50
100
150
200
250
300


Relative Error on Test Set = 0.85041
$X\theta$ (prediction)
y (true)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
50
100
150
200
250
300


$X\theta$ (prediction)
y (true)
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
3
4


$X\theta$ (prediction)
y (true)
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
3
4