Question 7

Question 1

For the objective of Question 2, Part 1, implement Heavy Ball. Try several momentum parameters and stepsizes, and find the best ones.

Using the data from Part 1, we implement heavy ball. Recall that the objective function F writes as

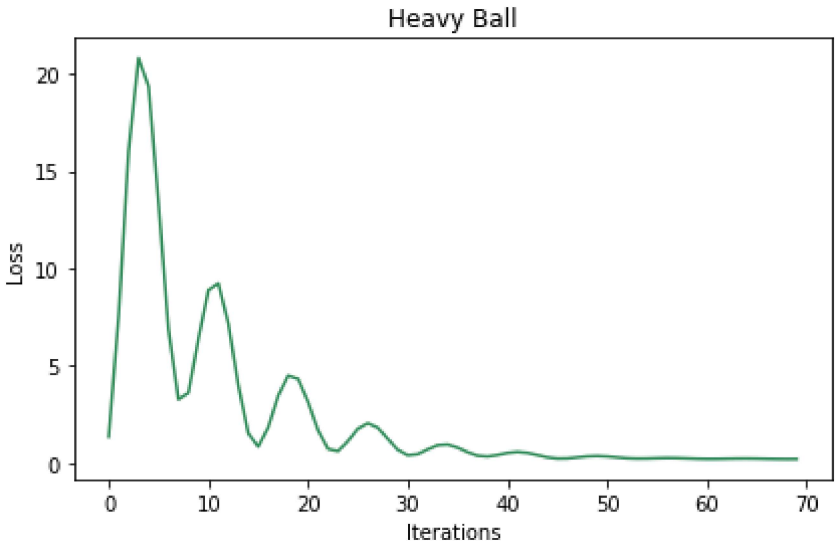
$$F : \theta \to \frac{1}{2n}||X\theta - y||_2^2,$$

with gradient equal to

$$\nabla F :\to \frac{1}{n} X^T(X\theta - y),$$

where $X \in \mathbb{R}^{n \times d}$, $\theta \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$

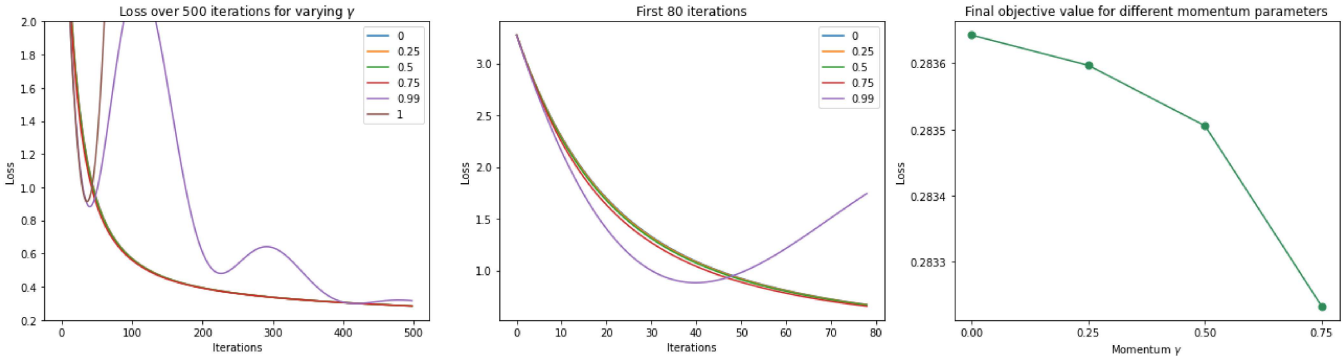
An initial run of 70 iterations with a step size $\alpha = 0.75$ and momentum parameter $\gamma = 0.9$ yields the following:



We observe the heavy ball method converging after an initial oscillating phase which corresponds well to the intuition of a ball rolling around the curvature of a bowl before finding the bottom.

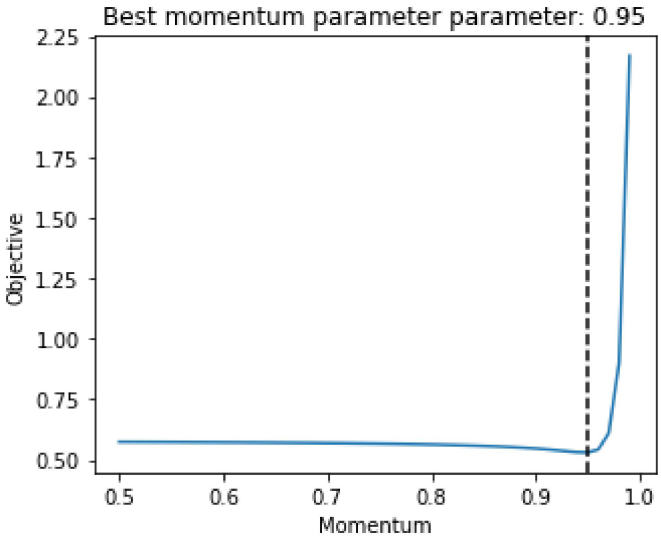
Momentum

We fix the step size at $\alpha = 0.014$ and for varying values of the momentum parameter γ perform the same procedure:

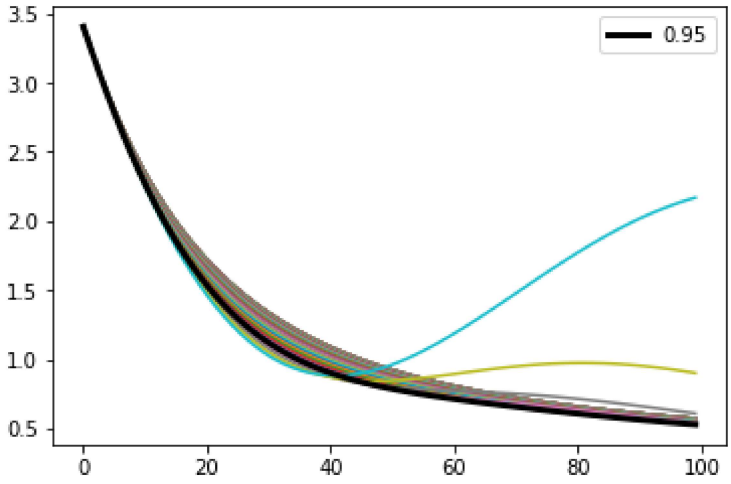


A momentum-parameter of 1 causes heavy-ball to diverge almost immediately and a parameter of 0.99 has the method oscillating before settling down. At first glance, the other parameters for momentum seem to have nearly exactly the same effect as no momentum at all in the convergence of the methods. We consider the other parameters and plot the final objective value over a very small run of 30 iterations since the method seems to converge so quickly.

As the method seems to converge quickly, we now run 100 iterationss each for 50 different values of γ and consider the final loss for each.



The optimal parameter for γ seems to be at 0.95. Indeed, the graph below shows the evolution of the error for a number of values, and we see that this momentum parameter moves the least.

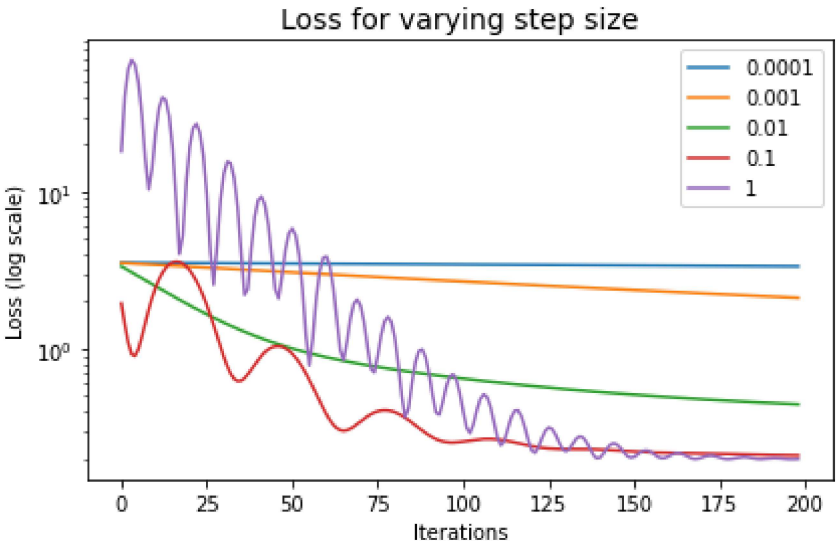


Step Size

Fix now the momentum parameter $\gamma = 0.95$, and for varying step sizes

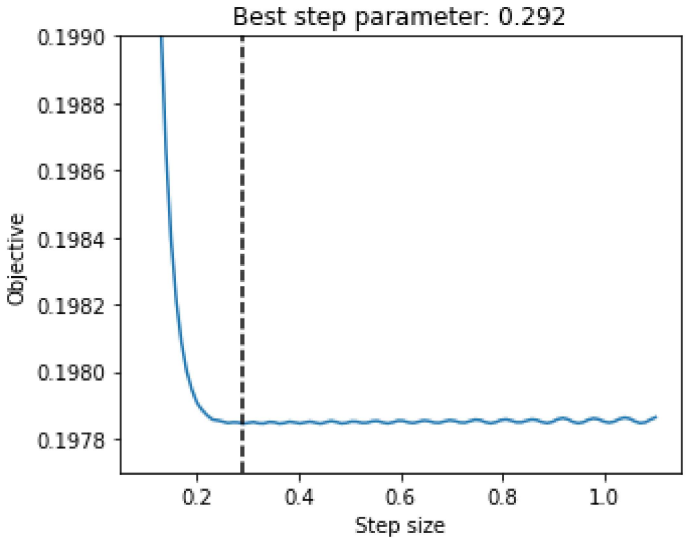
$$\alpha \in \{0.001, 0.01, 0.1, 1\},$$

we obtain the following:

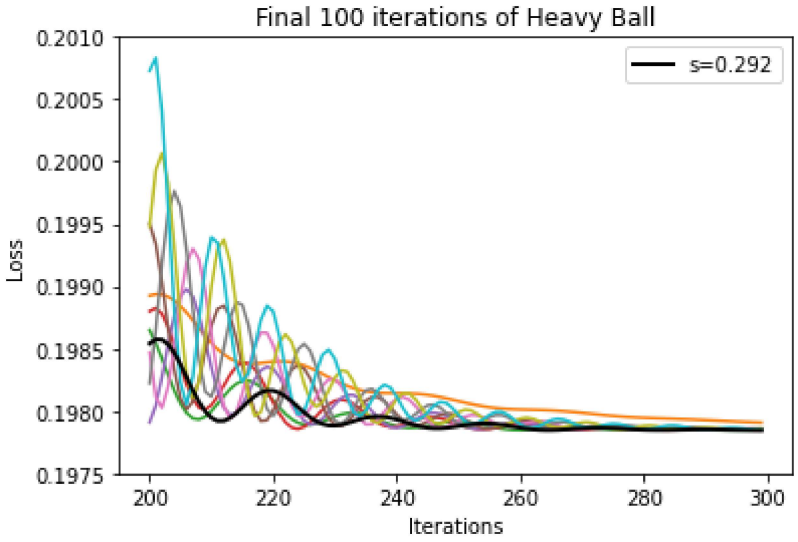


The above simulation seems to suggest that a step size of $\alpha = 0.1$ will provide the quickest convergence results. Step sizes that are too small seem to inhibit fast convergence.

To investigate this in greater detail, we run a simulation for 100 different values of the step parameter on the interval $[0.1, 1.1]$, and find the value that obtains the least loss for the training set.



The optimal step size, rounded to 3 significant digits, is 0.292. Indeed, we see for the final 200 iterations that the loss is still oscillating quite a bit, but that this step value is dying out fastest.

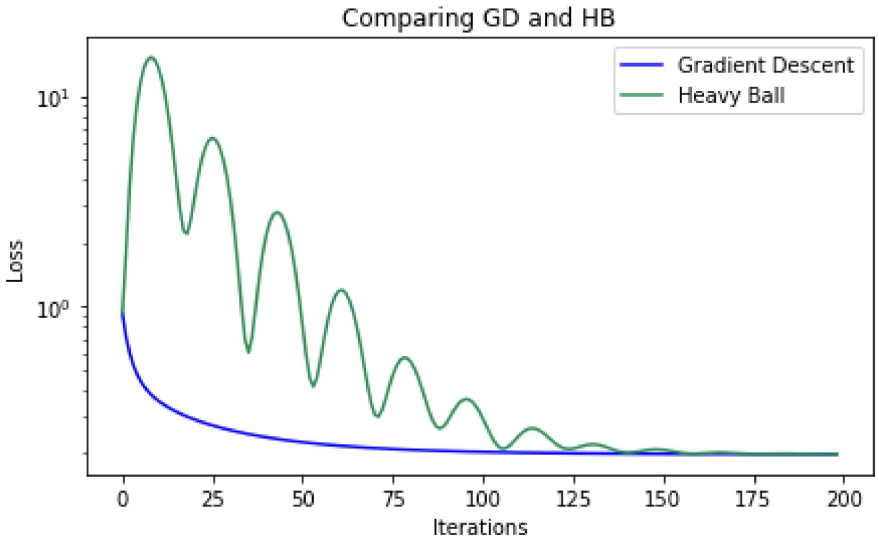


Question 2

With these parameters, how does Heavy Ball compare with gradient descent?

Recall that the best parameters are:

```
Step: 0.292
Gamma: 0.95
```



For the parameters tuned for heavy ball, we see heavy ball oscillating somewhat in the beginning but then settling out and converging after about 100 iterations. GD, on the other hand, converges much quicker but then stagnates after about 100 iterations.

Question 3: Non-convex optimisation

Propose a non-convex loss function for your regression or classification problem. Try minimizing the loss with gradient descent, and check that you reach at least an approximate first-order approximate critical point. Do you get good prediction results?

Consider now the objective function:

$$F : \theta \to \frac{1}{2n}||X\theta - y||_2^2 + \lambda||\theta||_p,$$

where $\theta \in \mathbb{R}^{n \times d}$ and $p \in (0, 1)$. The gradient writes as

$$\nabla F :\to \frac{1}{n} X^T(X\theta - y) + \lambda \nabla||\theta||_p$$

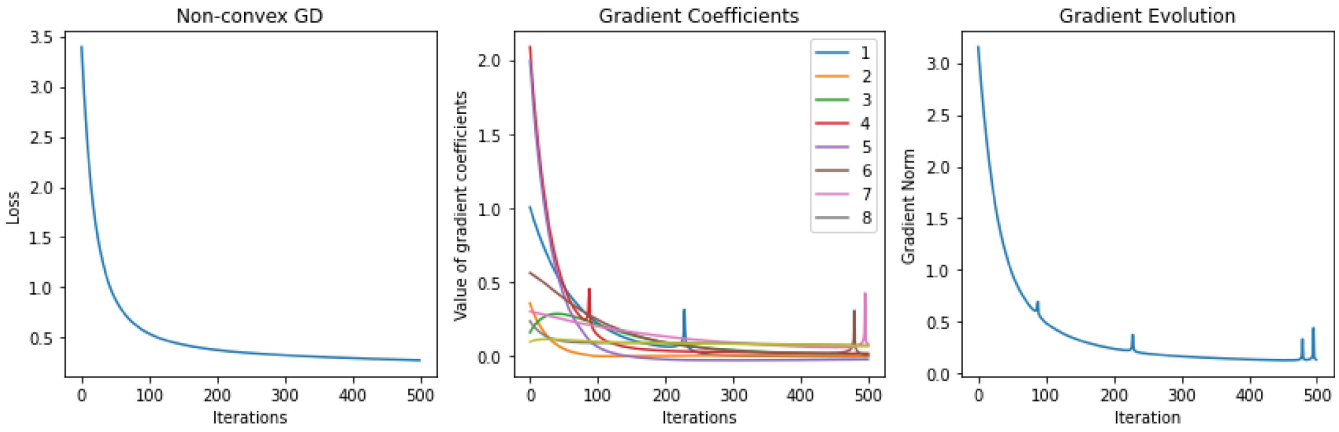
for $$ \nabla{||\theta||*p} = {||\theta||*{p}}^{1-p} \cdot [\theta_1^{p-1}, \theta_2^{p-1}, \dots, \theta_n^{p-1}]^T

We run 500 iterations of gradient descent with this non-convex regularisation term for the following parameters

- p=0.25
- $\alpha = 0.015$ (which is optimal for the original problem as determined in Part 1)
- $\lambda = 10^{-5}$

```
Running GD with a non-convex regulariser:

Final gradient coefficients:
[-0.00160573  0.00125058  0.01821395  0.01853524 -0.0211332   0.00919759
  0.07988003  0.07211536  0.06984595]
 Final gradient norm: 0.13293064018677092
```



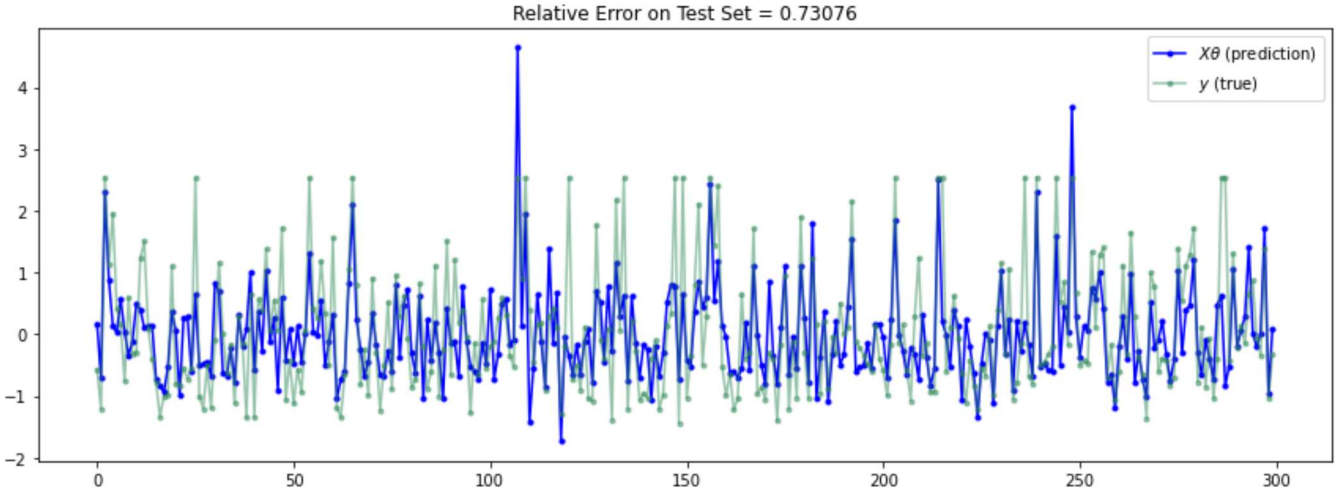
We have at least an approximate first-order critical point as the gradient at the last iteration has norm ~ 0.133.

Let us now check our ability to predict on the test set.

Recall that the model is specified by the very simple equation

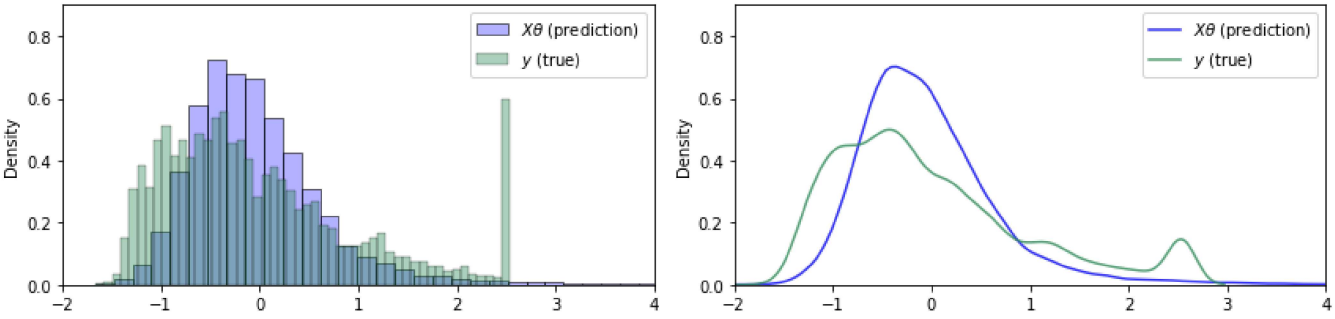
$$y = X\theta^*$$

and we have recovered $\hat{\theta}$ after running GD with a non-convex regulariser. We consider from the training set following: $\hat{y} = X_{train}\hat{\theta}$ and the true observed labels y^* to plot the differences in predictions and relative error.



The relative error on both the test and train set are comparable up to 2 significant digits and although not very close to 0, seems reasonable.

Consider also a histogram and kernel density plot of the observed and predicted lables y^* and $\hat{y}$ below:



We observe that these follow roughly the same distribution, indicating that long-term predictions would be statistically consistent.